TIETO- JA SÄHKÖTEKNIIKAN TIEDEKUNTA

Santtu Käpylä
Tiia Leinonen

PUHESYNTEESIN JA SUUN LIIKKEEN TOTEUTUS ROBOTILLE

Kandidaatintyö
Tietotekniikan tutkinto-ohjelma
Toukokuu 2020

Käpylä S., Leinonen T. (2020) Puhesynteesin ja suun liikkeen toteutus robotille. Oulun yliopisto, Tietotekniikan tutkinto-ohjelma, 44 s.

TIIVISTELMÄ

Ihmiset kommunikoivat keskenään sekä verbaalisesti, että nonverbaalisesti. Ihmisten verbaalisen viestinnän, eli puheen, ja ihmisäänen tutkimusta kutsutaan fonetiikaksi ja sitä hyödynnetään robotiikassa keinotekoisen puheen eli puhesynteesin toteuttamisessa. Sosiaalisten robottien yleistyessä on tärkeää, että ihminen-robotti-vuorovaikutus olisi mahdollisimman luontevaa. Varsinkin kasvot ovat avainasemassa vuorovaikutuksessa, ja robotin visuaalinen suu tekee vuorovaikutuksesta ihmiselle mielekkäämpää, sekä auttaa puheen ymmärtämisessä. Myös robotin puheen ja leuan liikkeen synkronointi on tärkeässä roolissa 'uncanny valley' -ilmiön välttämisessä.

Tässä projektissa toteutettiin puhesynteesi ja leuan liike InMoov-robotille. Leuan liike toteutettiin Dynamixel XL-320-servolla, jota kontrolloitiin Arduino UNO:lla. Puhesynteeseinä toteutettiin formanttisynteesi ja konkatenaatiosynteesi sekä mies-, että naisäänellä. Molempia puhesynteesejä sekä puheen ja suun liikkeen synkronisaatiota arvioitiin subjektiivisesti itse kehitetyillä numeerisilla asteikoilla. Konkatenaatiosynteesi saavutti arvosanan 3/5, mikä tarkoittaa, että puheen välittämä viesti on ymmärrettävissä, formanttisynteesi arvosanan 2/5, mikä tarkoittaa, että puheen välittämä viesti jää epäselväksi, ja synkronisaatio arvosanan 3/4, mikä tarkoittaa, että synkronisaatio on suurimman osan ajasta hyvä.

Projektissa hyödynnettiin apuna metakäyttöjärjestelmää Robot Operating System (ROS). Ohjelmistokomponenteista toteutettiin ROS-paketti, jonka toiminnallisuus koostuu kahdesta keskenään kommunikoivasta ROS-solmusta, joista toinen vastaa servon kontrolloimisesta ja toinen puhesynteesistä. Projektissa toteutettuja komponentteja voidaan hyödyntää tulevaisuudessa esimerkiksi ihminen-robotti-vuorovaikutuksen tutkimiseen ja erilaisten havaintoesitysten toteuttamiseen. Jatkossa puhesynteesin ja leuan liikkeen synkronisaatiota voitaisiin parantaa toteuttamalla monipuolisempia leuan asentoja sekä selkeyttämällä puhesynteesejä.

Avainsanat: robottipää, ihminen-robotti-vuorovaikutus, InMoov, ROS, konkatenaatiosynteesi, formanttisynteesi, Arduino, XL-320 servo

Käpylä S., Leinonen T. (2020) Speech Synthesis and Mouth Movement Implementation for Robot. University of Oulu, Degree Programme in Computer Science and Engineering, 44 p.

ABSTRACT

Humans communicate with each other both verbally and non-verbally. The research of verbal human communication and human voice is called phonetics and it can be utilized in robotics to produce artificial speech, ergo speech synthesis. As social robots become more common it is important for human-robot interaction to be as natural as possible. Especially faces are key components in interaction and robot's visual mouth makes the interaction more pleasant to humans and helps them in understanding speech better. The synchronisation of the robot's speech and mouth movement also plays an important role in avoiding the 'uncanny valley' phenomenon.

In this project two speech syntheses and motion of jaw were implemented for InMoov-robot. The motion of the jaw was implemented with Dynamixel's XL-320 servomotor controlled by Arduino UNO. Two different speech syntheses were implemented: formant synthesis and concatenation synthesis with both female and male voices. Both of the speech syntheses and the synchronization of the speech and mouth movement were evaluated on self-made subjective numerical scales. The concatenation synthesis accomplished a grade of 3/5, which means that the speech was comprehensible. The formant synthesis accomplished a grade of 2/5, which means the speech was quite incoherent. Last but not least, the synchronization accomplished a grade of 3/4, which means the synchronization was good most of the time.

The project utilizes a meta-operating system called Robot Operating System (ROS). The two software components implemented in this project comprise a ROS package consisting of two ROS nodes; one responsible for servo control and the other for speech synthesis. In the future these components can be utilized in research of human-robot interaction and in different demonstrations. The synchronisation of the jaw movement and speech could be improved by implementing more options for the position of the jaw and by enhancing the speech syntheses to sound more natural.

Keywords: robot head, human-robot interaction, HRI, InMoov, ROS, concatenation synthesis, formant synthesis, Arduino, XL-320 servomotor

SISÄLLYSLUETTELO

ALKULAUSE

Tämä kandidaatintyö on laadittu Oulun yliopiston tieto- ja sähkötekniikan tiedekunnassa kurssia ”Sulautettujen ohjelmistojen projekti” varten.

Oulussa 26. toukokuuta 2020

Santtu Käpylä
Tiia Leinonen

LYHENTEIDEN JA MERKKIEN SELITYKSET

AU	Toimintayksikkö, Action Unit
FACS	Facial Action Coding System
FS	Näytteenottoväli
F0	Perustaajuus
IPA	Kansainvälinen foneettinen aakkosto, International Phonetic Alphabet
ROS	Robot Operating System
TTS	Tekstistä puheeksi, Text-To-Speech

1. JOHDANTO

Puhe, eleet ja ilmeet muodostavat yhdessä ihmisten välisen luonnollisen kommunikaation perustan. Vuorovaikutustilanteessa ihmiset huomioivat puheesta sisällön lisäksi esimerkiksi äänen sävyä, korkeutta ja sanojen painotusta, joiden perusteella kuulija voi muodostaa mielikuvan puhujasta. Eleet ja ilmeet ovat myös tärkeitä ihmisten välisessä viestinnässä, sillä ne tuovat vuorovaikutukseen mukaan aivan oman ulottuvuutensa. Esimerkiksi tokaisu 'Onpa tänään kaunis päivä' saa aivan eri merkityksen, jos puhuja pyöritteleekin samalla silmiään sarkastisesti, ja erilaiset eleet auttavat elävöittämään puhujan viestiä entisestään.

Nykyään erilaiset koneet ovat monelle osa arkipäivää. Koneen käyttöliittymästä riippuen ihminen pystyy ohjaamaan konetta erilaisin syöttein ja vastaanottamaan koneelta tietoa esimerkiksi graafisesti sen ruudulta tai mittareiden lukemien avulla. Teollisuudessa on hyödynnetty tällaisia koneita ja robotteja jo pitkään hoitamaan erilaisia manuaalisia työtehtäviä automaattisesti ihmisten puolesta. Tekoäly ja robotiikka ovat viime vuosien aikana kuitenkin kehittyneet siihen pisteeseen, että robotteja voidaan alkaa käyttää myös yhteistyössä ihmisten kanssa, ja askelia siihen suuntaan on jo otettu. Tulevaisuudessa erilaiset sosiaaliset robotit voinevatkin toimia erilaisissa asiakaspalvelutöissä, kuten opettajina, ohjaajina tai hoitajina. Tätä varten ihmisen ja koneen välisen kommunikaation on muututtava vastaamaan ihmisten välistä luonnollista vuorovaikutusta.

Jotta sosiaaliset robotit voisivat toimia ihmisten parissa, on niillä oltava kyky kommunikoida kuten ihmiset. Puhesynteesin avulla sosiaaliset robotit saavat sananmukaisesti äänensä kuuluviin. On myös tärkeää, että robotin ulkoasu ja olemus ovat sellaisia, että ihmiset kokevat vuorovaikutuksen robotin kanssa miellyttäväksi ja helpoksi. Yleisesti voidaan ajatella, että mitä ihmisenkaltaisempi robotti on, sitä mielekkäämpää vuorovaikutus sen kanssa on. Toisaalta robotin lähestyessä liikaa ihmismäisyyttä ulkomuodoltaan ja eleiltään, voivat ihmiset alkaa kokea vuorovaikutuksen robotin kanssa epämiellyttäväksi.

Tässä projektissa kehitettiin InMoov-robotille leuan liikettä ja puhesynteesi, sekä näiden synkronisaatiota. InMoov-robotti on ensimmäinen avoimen lähdekoodin 3D-tulostettu ihmisen kokoinen humanoidirobotti. Tämä projekti on osa laajempaa kokonaisuutta, jossa pyritään toteuttamaan robotille erilaisia toiminnallisuuksia, jotta se kykenisi vuorovaikutukseen ihmisten kanssa. Yksinkertaisimmillaan käyttötilanne on sellainen, että robotti tervehtii ohikulkevia ihmisiä, ja houkuttelee heitä mukaan vuorovaikutustilanteeseen kanssaan.

2. MATKALLA IHMISEN JA ROBOTIN VUOROVAIKUTUKSEEN

Matka ihmisen ja robotin väliseen luonnolliseen vuorovaikutukseen on täynnä haasteita niin käytännön toteutuksen kuin eettisten kysymysten kannalta. Vaikka sosiaaliset robotit ovat kehittyneet paljon viimeisten vuosien saatossa, on täysin luonnollinen vuorovaikutus ihmisen ja robotin välillä toistaiseksi vielä tieteisfiktiota.

Luonnollinen kieli, jolla ihmiset kommunikoivat, koostuu sekä verbaalisesta, että nonverbaalisesta kommunikaatiosta [1]. Verbaalinen kommunikaatio, eli puhe tuotetaan puhe-elimillä [2]. Nonverbaalisessa kommunikaatiossa, eli esimerkiksi ilmeiden muodostamisessa varsinkin leuan liike on avainasemassa [3], ja leuan ja suun liikkeen muodostamiseen osallistuukin moninainen joukko kasvojen lihaksia [3, 4, 5].

Nykyisissä sosiaalisten robottien toteutuksissa verbaalinen kommunikointi toteutetaan puhesynteesillä, eli prosessilla, jolla tuotetaan keinotekoisesti ihmisääntä. Robottien ulkomuodot tukevat usein myös nonverbaalista kommunikointia mahdollistamalla ilmehtimisen. Varsinkin robotin visuaalinen suu tekee vuorovaikutuksesta ihmiselle mielekkäämpää [6] ja auttaa puheen ymmärtämisessä [7]. Suun toteutus on yleensä joko mekaaninen tai digitaalinen [8, 9].

Robottien on siis kyettävä toimimaan sosiaalisesti ja kommunikoimaan ihmisten tavoin, jotta ihmisten ja robottien välinen vuorovaikutus olisi mahdollisimman luontevaa [7, 8, 10]. Myös robottien käytön eettisyyttä tulisi tutkia eri ympäristöissä ja tehtävissä, sillä sosiaalisten robottien yleistyessä asiakaspalvelualoilla voidaan törmätä sekä moraalisiin että lainopillisiin kysymyksiin, kuten että voiko robottia asettaa vastuuseen ihmishengistä?

2.1. Ihmisten puhe, ilmeet ja suun anatomia

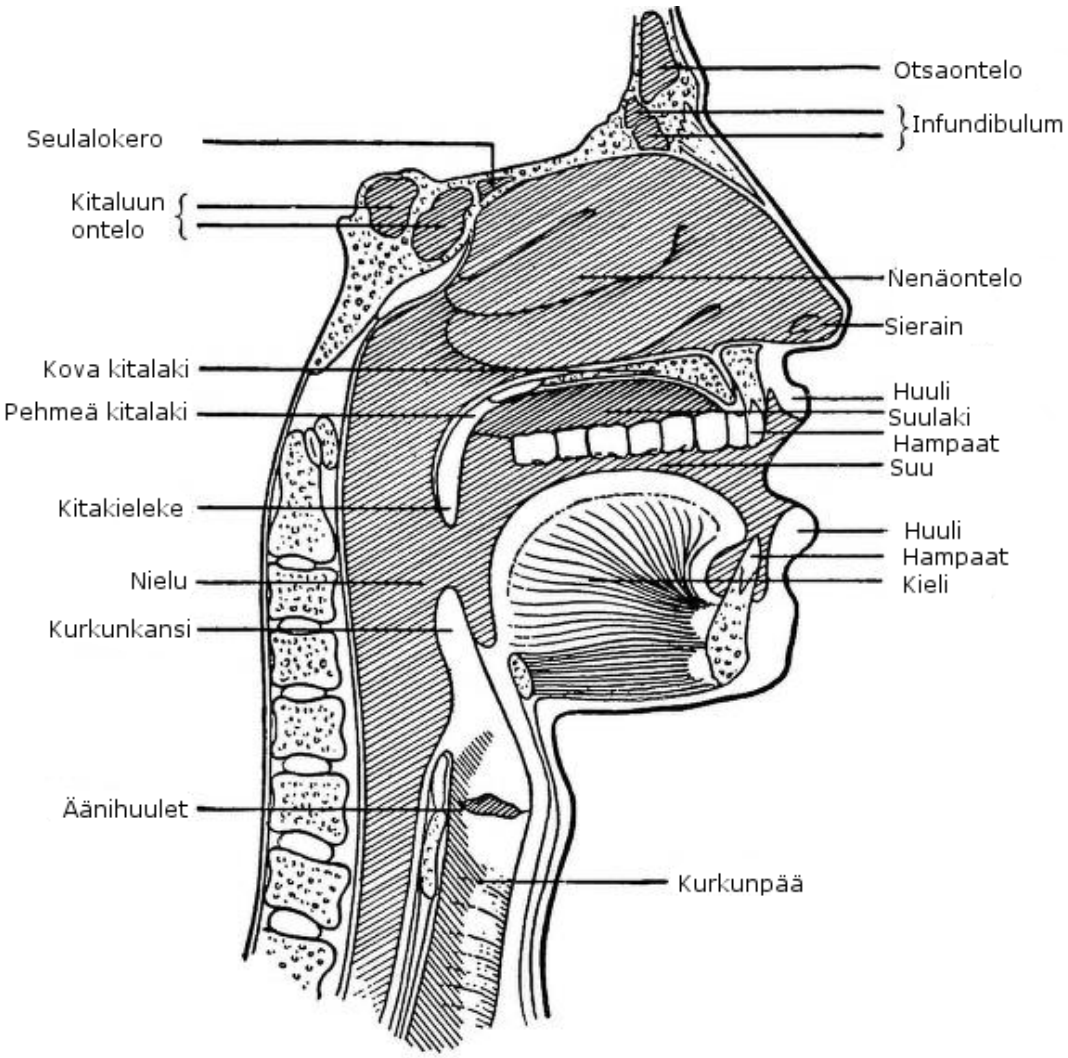
Ihmiset kommunikoivat keskenään puhumalla. Tätä verbaalista viestintää tutkitaan fonetiikan avulla. Fonetiikka on ihmisten puheen- ja äänentuoton tutkimusta, joka tarkastelee ihmisäänen syntyä, kulkua, sekä sen tunnistamista.

Monet kasvojen lihakset osallistuvat leuan ja suun liikkeiden synnyttämiseen [3, 4, 5]. Koska ihmiskasvojen rakenne on monimutkainen, on olemassa menetelmiä, joiden avulla voidaan helpottaa ilmeiden tuottamisen hahmottamista hyödyntämällä kartoitusta kasvojen eri piirteiden kytköksistä perustunteiden ilmaisuun [8]. Eräs tällainen menetelmä on nimeltään Facial Action Coding System (FACS) [11].

2.1.1. Fonetiikka

Fonetiikka käsittelee ihmisäänen syntyä, kulkua ja vastaanottamista. Ihmisäänen syntyä lähemmin tutkiva fonetiikan alalaji on artikulatorinen fonetiikka [12], jonka päämääränä on selvittää kuinka ääntöväylällä tuotetut äänet syntyvät ääntöväylällä olevien artikulaattoreiden avulla [12]. Äänet pystytään esittämään segmentteinä, jotka sisältävät vokaaleja ja konsonantteja [12]. Nämä segmentit sisältävät myös suprasegmentaalifoneemeja eli toissijaisia ominaisuuksia, joita ovat esimerkiksi painotus, intonaatio, pituus, sekä äänenkorkeus [2, 12, 13].

Suprasegmentaalifoneemien ominaisuuksien parametrisoinnilla ja tutkimisella voidaan tarkastella prosodiaa [2, 13], jossa olennaisina parametreinä ovat taajuus, intensiteetti ja kesto [2, 13].



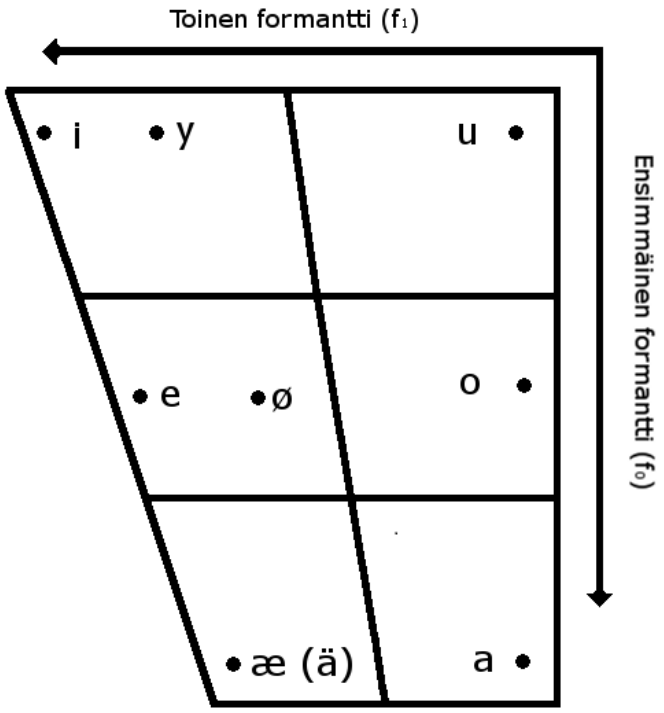
Kuva 1. Ihmisen puhe-elimet.[1]

Ihmiset käyttävät puhe-elimiä puheen tuottamiseen. Kuva 1 esittää ihmisen puhe-elimiä, joista tärkeimpiin kuuluvat mm. äänihuulet, nenäontelot, ääntöväylä, kieli ja huulet [2]. Puhe-elimet muodostavat yhtenäisen putken [2], jota pitkin keuhkoissa tuotettu ilmavirtaus kulkee puheen äänen tuottamiseksi [2, 13]. Vatsalihasten painaessa palleaa keuhkoista syntyy ulosmenevä ilmavirtaus, joka kulkeutuu henkitorveen ja kurkunpäähän [2, 13]. Kurkunpäässä sijaitsevan ääniraon jaksottaisella avaamisella ja sulkemisella äänihuulten avulla pystytään muuttamaan ilmavirtausta ja näin ollen tuottamaan ääniaaltoja [2, 13]. Näiden ääniaaltojen voidaan approksimoida olevan asymmetrisiä kolmioaaltoja [2]. Nielu yhdistää kurkunpään suu- ja nenäonteloihin, jotka muodostavat ääntöväylän [13]. Kurkunpään ja suuontelon tilavuutta ja mittasuhteita muuttamalla saadaan aikaan akustinen aikariippuvainen suodatin [13],

[1]Kuva muokattu teoksesta Fillebrown, Thomas (1836-1908): "Resonance in singing and speaking"[14], sivu 7. Tekijän kuolemasta yli 70 vuotta, joten kuva on tekijänoikeudeton.

jota käytetään eri äänteiden muodostamiseen. Suodatuksen jälkeen ääni kulkeutuu ympäristöön huulten ja sierainten kautta [13].

Jatkuva ääniaalto voidaan leikata pienemmiksi helposti tunnistettaviksi segmenteiksi, eli vokaaleiksi ja konsonanteiksi [12, 15]. Nämä segmentit voidaan yhdistää tavuiksi, jotka useimmiten sisältävät sekä vokaaleja, että konsonantteja, mutta joissain tavuissa on vain jompaakumpaa [15, 16], kuten tavuissa ’ei’ ja ’yö’. Suomen kielen natiivitavuissa on aina vokaali, ja yksittäiset konsonantit eivät voi muodostaa tavua [16], mutta esimerkiksi englannin kielessä tällaisia tavuja löytyy [15].



Kuva 2. Suomen kielen vokaalit ja niiden sijainnit ensimmäisen ja toisen formantin suhteen.[2]

Vokaalit tuotetaan äänihuulten värähtelemisen ja ääntöväylän resonaation avulla niin, että ääntöväylä on stabiilissa avoimessa asennossa ja siihen ei ole jäänyt ollenkaan ilmanpainetta [2]. Ääntöväylän resonanssitaajuuskaistaa kutsutaan formantiksi ja se karakterisoi jokaisen vokaalin [2, 12]. Kuvassa 2 on esitetty suomen kielen vokaalikaavio. Jokaisen äänteen formantin taajuus on eri, ja äänteen kahden ensimmäisen formantin katsotaan olevan tärkeimmät, sillä ne erottavat vokaaliäänteet toisistaan [2, 12]. Jos vokaali on tavun sisässä, se pystytään tarkemmin tunnistamaan viittaamalla sen fonologiseen asemaan kyseisessä tavussa [12]. Kahden eri vokaalin esiintymistä samassa äänteessä kutsutaan diftongiksi, ja yhden vokaalin esiintymistä monoftongiksi [16]. Suomen kielessä on kahdeksan vokaalia, jotka merkitään

[2]Kuvan tehnyt Santtu Käpylä. Kuvalle asetettu lisenssi CC BY 2.0 (`https://creativecommons.org/licenses/by/2.0/`).

kansainvälisen foneettisen aakkoston (International Phonetic Alphabet, IPA) mukaan seuraavasti: /a/, /e/, /i/, /o/, /u/, /y/, /æ/ ja /ø/ [16]. Missä /æ/ ja /ø/ ovat ä ja ö, sekä /o/ voidaan kirjoittaa joko o tai å. Muilla vokaaleilla kirjoitusasu pysyy samana suhteessa latinalaisiin aakkosiin.

Konsonantit tuotetaan ulosmenevällä ilmavirralla [12] ja ne lausutaan suhteellisen epävakaalla puhumisen konfiguraatiolla, varsinkin ääntöväylän muodon suhteen [2]. Ääntöväylä voi olla joko kokonaan tai osittain kiinni lyhyen aikaa konsonantista riippuen [2, 12]. Myös muut artikulaattorit, kuten kieli tai huulet, voivat aiheuttaa ilmavirran estymisen konsonantin tuottamiseksi [2, 12] ja sieraimien kautta ohjattu ilmavirta tuottaa nasaalikonsonantit [12]. Suomen kielessä on 17 konsonanttia, joiden IPA:n mukaiset esitystavat ovat: /b/, /d/, /f/, /g/, /h/, /j/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /t/, /ŋ/, /ʃ/ ja /ʋ/ [16]. Missä /ŋ/ on äng-äänne, /ʃ/ suhuässä ja /ʋ/ on w niiden diftongien jälkeen jotka päättyvät u:hun, kuten sanassa ’rouva’ (/rouʋa/) [16]. Joillain konsonanteilla on monia lausumismuotoja, vaikka ne merkitäänkin vain yhdellä IPA:n esitystavalla [16], esimerkiksi konsonantti /h/ lausutaan eri tavalla sanassa ’haamu’, kuin sanassa ’vaha’. Tällaisia konsonantteja ovat /p/, /t/, /k/, /s/, /h/, /m/, /n/, /l/, /r/, /ʋ/, ja /j/ [16].

Prosodian avulla voidaan tutkia puhutun kielen syntaksia, äännähdyksen mahdollista tyyppiä, sekä vuorovaikutuksen, asenteen ja tunteen ilmaisua [13]. Prosodisia ominaisuuksia tutkitaan usein lauseiden tai useiden sanojen yhteydessä [2], sillä pienemmissä äänen osissa, kuten tavuissa, ominaisuuden merkittävyyttä ei voi tutkia. Prosodisiin ominaisuuksiin kuuluvat mm. [2]:

- paino,
- äänenkorkeus,
- intonaatio,
- tauko,
- äänen voimakkuus,
- tempo ja
- parakieliset ominaisuudet.

Prosodian poikkeamien havaitseminen on helppoa [13], joten oikealla prosodian mallintamisella pystytään myös mallintamaan puheen lähdesignaalia [13], jota voidaan hyödyntää esimerkiksi luonnollisen puheen mallien luonnehtimisessa [13].

2.1.2. Suun ja leuan anatomia

Ihmisen leukojen liikuttamiseen osallistuu monitahoinen kokoonpano kasvojen lihaksia, joihin kuuluvat mm.:

- puremalihas (masseter),
- ohimolihas (temporalis),
- kaksirunkoinen alaleuan lihas (digastric),
- sisempi siipilihas (medialis pterygoid),
- ulompi siipilihas (lateralis pterygoid) ja
- kieliluun lihakset [3, 4, 5], sekä
- suun kehälihas (orbicularis oris) ja
- poskilihas (buccinator) [17].

Suun sulkemiseen, eli leuan nostamiseen osallistuvat edellämainituista lihaksista ohimolihas, puremalihas sekä sisempi siipilihas [4]. Suun avaamiseen, eli leuan laskemiseen osallistuvat puolestaan ulompi siipilihas, kaksirunkoinen alaleuan lihas sekä kieliluun lihakset [4].

Leuan eteenpäin suuntautuvasta liikkeestä vastaavat ulompi siipilihas, puremalihas ja sisempi siipilihas [4]. Taaksepäin suuntautuvasta liikkeestä vastaavat ohimolihas ja puremalihas [4]. Sivuttaisliikkeestä puolestaan vastaavat ohimolihas, puremalihas sekä molemmat siipilihakset [4].

Huulten liikkeestä vastaavat pääosin suun kehälihas, joka vastaa huulten sulkemisesta ja eteenpäin suuntautuvasta liikkeestä, sekä poskilihas, jonka tehtävä on vetää suupieliä ylöspäin [17].

Taulukko 1. Suun ja leuan liikkeet ja niistä vastaavat lihakset.

Lihas	Liike						
	Leuka					Huulet suppuun	Suupielet ylös
	ylös	alas	eteen	taakse	sivulle		
puremalihas	x		x	x	x		
ohimolihas	x			x	x		
kaksirunkoinen alaleuan lihas		x					
sisempi siipilihas	x		x		x		
ulompi siipilihas		x	x		x		
kieliluun lihakset		x					
suun kehälihas						x	
poskilihas							x

Taulukkoon 1 on koostettu erilaisia suun liikkeitä ja kyseisiin liikkeisiin osallistuvat lihakset. Tietoa ja ymmärrystä kasvojen lihasten toiminnasta voidaan hyödyntää esmierkiksi ihmisten kasvojen ilmeiden muodostamisen tutkimisessa, sekä realististen robottikasvojen kehittämisessä [18].

2.1.3. Facial Action Coding System

Koska ihmisen kasvoissa on edellämainittujen suun liikkeistä vastaavien lihasten lisäksi yhteensä yli 40 lihasta, jotka osallistuvat ilmeiden tuottamiseen [11], on ihmisen kasvojen anatomiaa mukailevan robottikasvojen toteutuksen kontrollointi käytännössä osoittautunut vaikeaksi [3, 9, 11, 18]. Jotta ilmeiden tuottamiseen osallistuvien lihaskokonaisuuksien hahmottaminen olisi yksinkertaisempaa, voidaan käyttää menetelmiä, jotka hyödyntävät kartoitusta kasvojen eri piirteiden kytköksistä perustunteiden ilmaisuun [8]. Tarkasta perustunteiden määrästä ei olla yksimielisiä, mutta merkittävimpiä teorioita voi tarkastella mm. asialle omistautuneilta verkkosivuilta [19]. Esimerkiksi Ekmanin teorian mukaan perustunteiden ajatellaan olevan viha, inho, pelko, ilo, suru ja yllättyneisyys [19].

FACS (Facial Action Coding System) on menetelmä, jonka avulla mikä tahansa kasvojen ilme on mahdollista koodata ns. toimintayksiköiden (action unit, AU) arvojen avulla [11]. Tietty kasvojen ilme voidaan koostaa aktivoimalla jokainen siihen osallistuva AU ja antamalla niille jokaiselle intensiteettiarvo [11]. Suun alueen merkittäviä AU:ja ovat esimerkiksi suupielten laskeminen, huulten supistaminen, leuan nosto sekä huulten venytys [11].

2.2. Robottien puhesynteesi ja keinotekoiset suut

Jotta robottien ja ihmisten vuorovaikutus olisi mielekästä, on robottien kyettävä kommunikoimaan verbaalisesti ja nonverbaalisesti ihmisten tavoin [8, 7, 10].

Verbaalinen kommunikointi toteutetaan puhesynteesillä, eli prosessilla, jolla tuotetaan keinotekoisesti ihmisääntä. Synteesi voidaan toteuttaa digitaalisesti algoritmeilla, sähköisesti komponenteilla, tai fyysisillä mekanismeilla [20]. Puhesynteesin avulla tuotettua ääntä voidaan hyödyntää ihmisen ja robotin välisen suullisen kommunikaation toteuttamisessa.

Sosiaalisilla roboteilla on usein myös visuaalinen suu, jotta nonverbaalinen vuorovaikutus olisi mielekkäämpää ihmisen kannalta [6]. Suun toteutustapa on yleensä digitaalinen tai mekaaninen, riippuen kasvojen toteutustavasta [9]. Huulten liikkeiden ja puheen synkronisaatiolla on suuri merkitys siihen, kuinka realistiseksi robotti mielletään [7].

2.2.1. Puhesynteesi

Viimeisten kahden vuosisadan aikana foneettisten ilmiöiden tutkimiseen on kehitetty ja käytetty laitteita, jotka voivat tuottaa synteettistä puhetta [2, 13, 20]. Ensimmäiset puhesynteesijärjestelmät olivat fyysisiä mekanismeja, jotka simuloivat ihmisen kurkunpäätä ja ääntöväylää [2, 13, 20]. Nämä järjestelmät demonstroivat, että puheen tuottaminen voidaan reaalisoida fyysisillä mekanismeilla, mikä oli aikaisemmin vain teoreettinen konsepti [20]. Mekaanisten puhesynteesijärjestelmien kohtuullinen menestys antoi toivoa, että riittävällä kontrollilla tällaisilla laitteilla voitaisiin tuottaa ymmärrettävää puhetta [20], mikä antaisi esimerkiksi puhevikaisille henkilöille apuvälineen keskusteluun.

1900-luvun puolivälissä synteettisen puheen laitteistot muuttuivat suuresti sähköisten järjestelmien saapumisen myötä [2, 20], sillä mekaanisten osien korvaaminen virtapiireillä mahdollisti monimutkaisempia ja joustavampia toteutuksia puheen tuottamiseen [20]. Digitaaliteknologia mullisti synteettisen puheen muodostamisen, koska fyysiset laitteet pystyttiin nyt toteuttamaan algoritmien avulla [20]. Vaikka monet digitaaliset puhesynteesijärjestelmät ovat periaatteessa laskennallisia versioita sähköisistä järjestelmistä, algoritmit parantavat ja lisäävät merkittävästi uusia tapoja toteuttaa synteettistä puhetta [20]. Lisäksi toteutusten testaaminen ja muokkaus on helpompaa, sillä algoritmin koodin muokkaaminen on huomattavasti helpompaa, kuin virtapiirin uudelleensuunnittelu.

Ihmisen ääntöväylän mekaniikan ja sen tuottaman puhesignaalin mallintamista kutsutaan artikulatoriseksi synteesiksi [2, 13, 20, 21, 22]. Koska artikulatorinen

synteesi yrittää mallintaa luonnollista puheen tuottamista mahdollisimman tarkasti, se on teoreettisesti paras menetelmä korkealaatuiseen puhesynteesin tuottamiseksi [13]. Vaikea käytännön toteutus ja raskas laskennallinen kuorma kuitenkin tuottavat ongelmia [13]. Nykyisten puheentuottomallien ja laskentatehon rajoitteellisuuden takia artikulatorinen puhesynteesi ei ole menestynyt yhtä hyvin kuin muut puhesynteesimenetelmät [13]. Tulevaisuudessa tämä voi muuttua, sillä parempia artikulaatiomalleja on kehitteillä ja laskentatehon resurssit ovat nousussa [13].

Tekstistä puheeksi (Text-to-speech, TTS) on prosessi, joka muuttaa syötteeksi annetun tekstin ääniaaltomuodoksi [2, 13, 23]. TTS-järjestelmällä on kaksi pääosaa: tekstin analyysi ja puheaaltomuodon tuottaminen [2, 13, 23]. Tekstin analyysissä syöte muutetaan foneettiseen tai muuhun kielelliseen esitystapaan ja sen prosodiaa arvioidaan [2, 13, 23]. Kielitieteellisellä analyysilla tekstistä voidaan hahmotella sanamuodot ja sanojen painomallit [13]. Analyysin ja rakenteellisen tiedon avulla statistisia menetelmiä hyödyntäen voidaan päätellä äänen taajuuden korkeuskäyrä, sekä äänteiden kesto [13]. Yksinkertaisimmillaan näillä tiedoilla voidaan toteuttaa toinen pääosista, puheaaltomuodon tuottaminen [13]. Jotkin toteutukset yrittävät jättää tekstin analyysin välistä, toteuttaen end-to-end TTS:n, missä analyysiä on esimerkiksi yritetty vähentää harjoittamalla syntetisoijaa teksti- ja äänipareilla [24]. Nykyiset vaihtoehdot ovat jo merkittävästi vähentäneet analyysien määrää, mutta eivät poistaneet niitä kokonaan [24, 25]. Esimerkkejä end-to-end TTS:istä ovat Tacotron [24] ja Char2Wav [25].

Formanttisynteesi nimensä mukaan mallintaa puhumisen formantteja [20, 21]. Yleisiä parametreja tällaiselle synteesille ovat taajuudet ja amplitudit resonanssille, sekä perustaajuus [20]. Esimerkiksi sana ’juu’ voidaan tuottaa yksinkertaisella toisen formantin interpoloinnilla korkeasta taajuudesta paljon matalampaan taajuuteen, kuten 2200 Hz:stä 400 Hz:iin, muuttamatta muita parametrejä [20]. Interpoloitua taajuutta käytetään sen jälkeen syntetisoijan asetusten muuttamiseen tietyn ajanjakson yli, muodostaen ääniaallon, joka muistuttaa sanaa ’juu’ [20].

Nauhoitettujen puheäänien yhdistämistä puhesignaaliksi kutsutaan konkatenaatiosynteesiksi [13, 20]. Yhdistäminen toteutetaan etsimällä lyhyitä äänteitä ja ääniparej, eli difoneja algoritmien avulla, ja asettamalla löydetyt äänteet peräkkäin aikatasoon, luoden synteettistä puhetta [20]. Konkatenaatiosynteesillä tuotettu puhe voidaan syntetisoida hyvin ymmerrettäväksi ja luonnolliseksi [13], sillä ihmisäänen luonnolliset tunnusmerkit säilyvät nauhoitettuihin äänteisiin [13]. Äänteiden yhdistämisessä voi kuitenkin tapahtua vääristymisiä nauhoitettujen äänteiden jatkumattomuuksien takia [13], jonka seurauksena äänen sulavuus on toteutuksissa yleensä matala. Toisaalta, esimerkiksi Amazonin Alexa on malliesimerkki sulavasta konkatenaatiosynteesistä. Nauhoitettujen äänien määrä on aina rajoitettu, mikä tekee konkatenaatiosynteesin puheesta vähemmän joustavaa, sillä vain yhden puhujan äänenlaatua voidaan imitoida [13]. Nykyään konkatenaatiosynteesi on ehkä yksi yleisimmista puhesynteesin muodoista sen luonnollisuuden takia [13], vaikka parantamisen varaa löytyy.

Tilastollisessa parametrisynteesissä puheen samanlaisista parametreistä otetaan estimaatit, joista muodostetaan uudet käytettävät segmentit, jotka ovat vähemmän luonnollisia kuin konkatenaatiosynteesissä käytetyt alkuperäiset nauhoitetut äänet [2, 26], mutta puheen tyyliä, tuntomerkkejä ja tunnesisältöä voidaan muuttaa parametrien muokkauksesta riippuen [13]. Parametrit muodostetaan yleensä kielitieteellisten,

prosodisten, ja foneettisten ominaisuuksien mukaan [26]. Parametreinä voivat toimia esimerkiksi [26]:

- ensimmäinen ja toinen formantti;
- nykyinen, edeltävä, ja seuraava foneemi, sekä sen paikka tavussa;
- tavun sisältämät foneemit ja niiden määrä nykyisissä, edeltävissä ja tulevissa tavuissa;
- tavujen paino ja korostus, sekä etäisyys edeltävään ja tulevaan painoon tai korostukseen;
- vokaalin identiteetti nykyisessä tavussa;
- sanojen määrä, niiden edeltäjät ja seuraajat, sekä sanojen sisältämien tavujen määrä;
- sisältösanojen määrä ja paikat verrattuna toisiinsa; ja
- koko lausunnon sisältämät tavu-, sana- ja virkemäärät.

Valituista parametreistä muodostetaan estimaatit, ja estimointiin voi esimerkiksi käyttää suurimman uskottavuuden menetelmää [26] tai lineaarista ennustavaa koodausta [2, 13]. Usein myös Markovin piilomalleja käytetään tilastollisessa parametrisynteesissä [13, 26]. Taulukkoon 2 on koostettu eri puhesynteesien vertailua, vertailun tulokset ovat keskiarvoisia tuloksia, eli kaikkien konkatenaatiosynteesien äänensulavuus ei ole matala, mutta näin on useimmissa tapauksissa.

Taulukko 2. Puhesynteesien vertailua

	Konkatenaatio	Formantti	Tilastollinen Parametri
Luonnollisuus	korkea	matala	keskiverto
Äänen sulavuus	matala	keskiverto/korkea	korkea
Sanasto	keskiverto/korkea	loputon	korkea
Toteutuksen vaikeus	helppo	helppo	valitusta mallista riippuva

2.2.2. Robottien suut

Robotin suun toteutus riippuu paljon robotin käyttötarkoituksesta ja kohderyhmästä, jonka parissa sen on tarkoitus toimia. Suut voivat olla realistisia tai karikatyyrisia [8], tai toisinaan suulle ei ole tarvetta ollenkaan [6]. Robottikasvojen, ja siten samalla myös robottisuiden yleisimmät toteutustavat ovat mekaaninen ja digitaalinen [8, 9]. Myös ns. hybriditoteutus, eli sekoitus molempia on mahdollinen [9]. Suun ulkonäöllä ja liikkeillä on vaikutusta mm. siihen, kuinka ystävälliseksi tai turvalliseksi robotti mielletään [1, 7, 6].

Suuttomasta robotista esimerkkinä toimii Keepon (Kuva 3), robotti joka on suunniteltu nonverbaaliseen kanssakäymiseen lasten kanssa [27]. Robotin ulkoasu on tarkoituksella minimalistinen, jotta vuorovaikutus sen kanssa olisi mahdollisimman luonnollista ja intuitiivista lapsille [27]. Sunniteltaessa ihmisenkaltaista robottia

suuttomuus voi kuitenkin johtaa siihen, että ihminen ei miellä robottia elollisen kaltaiseksi olennoksi [6]. Suun puuttuminen lisää myös riskiä, että robotin tunnetilat tunnistetaan väärin; suutonta robottia pidetään helpommin tahattomasti surullisena [6].

Kuva 3. Keepon, suuton robotti.[3]

Useimmiten on siis tarkoituksenmukaista, että robotilla on edes jonkinlainen suu, vaikka se olisikin pelkkä staattinen suu. Tästä hyvä esimerkki on Pepper (Kuva 4a). Pepper on ulkomuodoltaan suunniteltu androgyynin ihmismäiseksi, mutta sen fyysinen, staattinen suu ja suuret silmät auttavat välttämään 'uncanny valley'-ilmiön syntymisen [28]. Staattisella suulla ei kuitenkaan saavuteta yhtä hyvää ihmisenkaltaisuutta kuin liikkuvalla suulla [6].

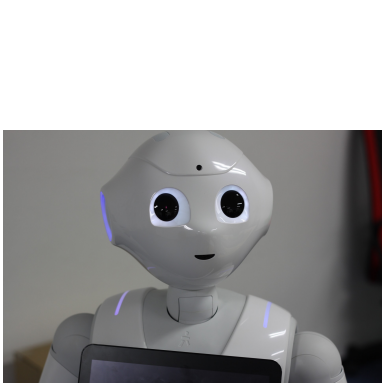
Yksinkertaisesta mekaanisesta suun toteutuksesta esimerkkinä toimii FLASH [6, 10] (Kuva 4c). FLASH:in pää koostuu kolmesta levystä, joista alin, pystysuunnassa liikuteltava levy, toimii robotin leukana [10]. FLASH:in suun toteutus on kuitenkin aika rajoittunut, sillä osa sen ilmaisemista tunteista on ihmiselle hankalaa tunnistaa suun toteutuksen yksinkertaisuuden vuoksi [10]. Suun liikkeen ja äänen synkronisaatio pelkällä suun avaamis- ja sulkemisliikkeellä on myös erittäin epärealistinen toteutus [7].

Monimutkaisemmasta mekaanisesta suun toteutuksesta malliesimerkki on Kismet (Kuva 4b). Sillä on 5-vapausasteinen suu, joista neljä on huulia ja yksi leukaa varten [7]. Kismet kykenee kuuteen eri suun asentoon, jotka ilmentävät inhoa, pelkoa, surua, yllätystä, vihaa ja iloa [29] Ekmanin perustunteiden teorian mukaan [19]. Kismetille on myös toteutettu huulten liikkeen ja puheen synkronisointia [29]. Kismetin visuaalinen toteutus on kuitenkin karikatyyrinen, joten sekä sen suun tuottamien ilmeiden, että huulten liikkeen ja puheen synkronisaation toteutus on tarkoituksella jätetty yksinkertaiseksi, sillä äärimmäisen realistinen toteutus voisi näyttää pakotetulta sarjakuvamaisella hahmolla [29].

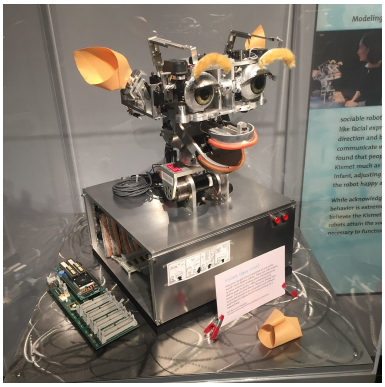
Kaikilta edellämainituilta robottikasvoilta puuttuu kuitenkin iho, mikä vähentää niiden ihmismäisyyttä. Albert HUBO -robotilla (Kuva 4d) on 'Frubber'-nimisestä materiaalista tehty keinotekoinen iho, joka tekee robotin kasvoista ihmismäiset [30]. Kasvojen ilmeet tuotetaan servomoottoreilla, joita kontrolloimalla voidaan kiristää ja

[3]Kuva 'IMG_0253' tekijältä paulbettner (https://www.flickr.com/photos/72791776@N00), saatu lisenssin CC BY 2.0 alaisena (https://creativecommons.org/licenses/by/2.0/).

löysätä kasvojen eri kohtiin, kuten huuliin ja leukaan, kiinnitettyjä naruja [30]. Näin voidaan matkia ihmiskasvojen luonnollisia ilmeitä.



(a) Pepper.



(b) Kismet.

(c) FLASH.

(d) Albert HUBO.

Kuva 4. Esimerkkejä mekaanisin suin varustetuista roboteista.[4]

Mekaanisten keinojen lisäksi robottisuiden toteutustavat voivat olla digitaalisia. Yleisimmät keinot digitaalisten suiden esittämiseen ovat erityyppisten valojen ja näyttöjen käyttö. Digitaalisuuden etuja ovat muun muassa helpompi toteutus sekä muunneltavuus [8].

[4]Pepperin kuva käyttäjältä imjanuary (https://pixabay.com/users/imjanuary-2490745/), saatu Pixabay-lisenssin alaisena (https://pixabay.com/service/terms/#license).

Kuva 'Kismet' käyttäjältä sillygwailo (http://www.flickr.com/photos/35034348378@N01), saatu lisenssin CC BY 2.0 alaisena (https://creativecommons.org/licenses/by/2.0/).

FLASH:in kuva 'L1007567' käyttäjältä foam (https://www.flickr.com/people/foam/), saatu lisenssin CC BY-SA 2.0 alaisena (https://creativecommons.org/licenses/by-sa/2.0/).

Kuva 'Einstein-Hubo' käyttäjältä Dayofid (https://en.wikipedia.org/wiki/User:Dayofid), saatu lisenssin CC BY 2.5 alaisena (https://creativecommons.org/licenses/by/2.5/).

(a) Buddy

(b) Snackbot

(c) Furhat

Kuva 5. Esimerkkejä erilaisin digitaalisin suin varustetuista roboteista.[5]

Yksinkertaisimmillaan digitaalisen suun toteutus on vain valo, joka välkkyy päälle ja pois robotin puheen mukana [6]. Useimmin valoja on kuitenkin enemmän kuin yksi, jolloin toteutus monipuolistuu. Robottisuiden LED-toteutuksesta esimerkkinä toimii Snackbot (Kuva 5b). Snackbotin suuna toimii aivan sen kasvojen alaosassa sijaitseva 3x12 LED-näyttö, jolle on ohjelmoitu erilaisia suun animaatioita, esimerkiksi eri suun muotoja, liikkeitä ja värejä [31]. Monimutkaisemmissa LED-toteutuksessa voidaan hyödyntää myös LED-matriiseja, joiden avulla voidaan suun lisäksi esittää myös muita muotoja, ja jopa tekstiviestejä [6].

Robotin kasvot, ja siten myös suut, voidaan myös toteuttaa 2D-näytöillä animoidusti. Esimerkiksi Buddy-robotin (Kuva 5a) kasvot koostuvat animoiduista

[5]Buddyn kuva käyttäjältä Pierre Metivier (https://www.flickr.com/people/feuilllu/), saatu lisenssin CC BY-SA 2.0 alaisena (https://creativecommons.org/licenses/by-nc/2.0/).

Snackbotin kuva käyttäjältä Jiuguang Wang (https://www.flickr.com/people/jiuguangw/), saatu lisenssin CC BY-SA 2.0 alaisena (https://creativecommons.org/licenses/by-sa/2.0/).

Furhatin kuva käyttäjältä Rain Rabbit (https://www.flickr.com/people/37996583811@N01/), saatu lisenssin CC BY-NC 2.0 alaisena (https://creativecommons.org/licenses/by-nc/2.0/).

silmistä ja suusta, joiden muodostamia ilmekokonaisuuksia voidaan esittää kosketusnäytöltä [6, 32]. Kaksiulotteiset näytöt kärsivät kuitenkin ns. Mona Lisa -efektistä, eli siitä, että kolmiulotteisen objektin orientaatio suhteessa tarkkailijaan mielletään vakioksi, riippumatta siitä mistä suunnasta tarkkailija objektia katsoo [9], eli esimerkiksi tapauksessa jossa objektina ovat animoidut kasvot, tarkkailija joko muodostaa katsekontaktin kasvojen kanssa, kuten Mona Lisa -maalauksen tapauksessa, tai kokee että kasvot 'katsovat' tiettyyn suuntaan, riippumatta siitä missä kohtaa tarkkailija seisoo suhteessa kasvoihin.

Kun halutaan luoda mahdollisimman ihmismäinen robotti, voidaan hyödyntää useita eri toteutustapoja. Esimerkiksi Furhat (Kuva 5c) on ns. hybriditoteutus, eli robottipää jossa kasvoissa hyödynnetään sekä digitaalista että fyysistä toteutustapaa: robotilla on 3D-mallinnetut kasvot, joille projektoidaan pään takapuolelta kasvojen liikkeen toteuttava animaatio [9]. Käyttämällä fyysistä kasvojen mallia kasvojen perustana, kasvoista saadaan luonnollisemman oloiset kuin tasaisella näytöllä esitetyistä kasvoista, ja vältetään Mona Lisa -efektin syntyminen [9]. Kasvojen digitaalisen toteutuksen avulla puolestaan vältytään kasvojen liikkeiden fyysisen toteutuksen haasteilta [8, 9].

2.3. Ihmisen ja robotin välinen vuorovaikutus

Yleisesti voidaan ajatella, että mitä enemmän robotti on ihmisenkaltainen, sitä miellyttävämpää ihminen-robotti-vuorovaikutus on [33]. Kuitenkin robotin ollessa liian ihmismäinen, ihminen voi alkaa kokea vuorovaikutuksen epämiellyttäväksi, mitä kutsutaan 'uncanny valley' -ilmiöksi [8, 33, 34].

Ihmisen ja robotin vuorovaikutuksen kannalta kasvot ovat tärkeässä roolissa [8, 10], sillä eleet ja ilmeet auttavat välittämään viestin tunnesisällön vastaanottajalle [8, 33]. Myös robotin puheen toteutustavalla on vaikutusta siihen, millaiseksi vuorovaikutus robotin kanssa mielletään [35, 36].

Tulevaisuudessa sosiaalisten robottien yleistyessä, on kuitenkin tärkeää pysähtyä miettimään sosiaalisten robottien käytön eettisyyttä ja lainopillista puolta eri aloilla ja erilaisten kohderyhmien parissa [34].

2.3.1. Uncanny valley -ilmiö

Yleisesti, mitä realistisempi robotti on, sitä positiivisempaa sen kanssa käyty ihminen-robotti-vuorovaikutus on [33]. Kuitenkin, jos robotin toiminnallisuus ei vastaa ihmisen odotuksia, esimerkiksi suun liike ja puhe eivät ole sopusoinnussa, ihminen kokee vuorovaikutuksen epämiellyttävänä [8]. Tätä ilmiötä, jossa ihminen kokee robotin olevan ihmismäinen, mutta vääristyneellä tai elottomalla tavalla, kutsutaan nimellä 'uncanny valley', eli vapaasti suomennettuna 'outo laakso' [8, 33, 34].

Antropomorfisilla, eli ihmistä muistuttavilla toteutuksilla 'uncanny valley' -ilmiön syntyminen on sitä todennäköisempää, mitä realistisempi toteutus on kyseessä [1, 33], sillä liiallinen realismi voi kiinnittää huomion ihmisen ja robotin eriäväisyyksiin yhteneväisyyksien sijaan [1]. Zoomorfisilla, eli eläintä muistuttavilla toteutuksilla 'uncanny valley' -ilmiön syntyminen on epätodennäköisempää [1]. Karikatyyrisillä

kasvojen toteutuksilla ei ehkä saavuteta realistista lopputulosta, mutta riski ’uncanny valley’-ilmiön syntyyn on pienempi [1].

Robottien suunnittelijoiden kannattaa siis panostaa yksityiskohtiin ihmisenkaltaisen robotin toteutusvaiheessa, tai tarkoituksella pitäytyä vähemmän antropomorfisissa toteutuksissa, jolloin ’uncanny valley’ -vaikutusta ei synny [37].

2.3.2. Kasvojen merkitys vuorovaikutuksessa

Ihmisten kannalta kasvot [8, 10], ja varsinkin leuat [3] ja suut [6] ovat tärkeitä vuorovaikutuksessa; puhujan huulten liike auttaa kuulijaa viestin ymmärtämisessä ja puhujan identifioimisessa ihmisjoukosta [6]. On myös osoitettu, että suurin osa viestin tunnesisällöstä välittyy vastaanottajalle nonverbaalisesti eleiden, ilmeiden ja kasvojen liikkeen välityksellä [8, 33]. Siksi kasvojen ilmeet ja eleet ovat tärkeässä roolissa ihminen-robotti-vuorovaikutuksessa [3, 30]. Kasvojen välittämän tunnetason intensiteetti riippuu kasvojen realistisuudesta [33]. Siksi on ehdotettu, että ihmismäisten robottikasvojen kannattaisi hieman liioitella ihmisen ilmeitä saavuttaakseen saman intensiivisyystason robottikasvojen ilmeille [33]. Robottien kasvojen ja äänen yhteensopivuus auttaa tekemään robotista uskottavamman oloisen [34]. Epäsynkroninen suun liike voi myös vaikeuttaa puheen tunnistamista, sillä kuullessaan äänen joka ei sovi yhteen näköhavainnon kanssa, kuulija voi mennä sekaisin äänen tulkinnassa [7].

2.3.3. Puheen merkitys vuorovaikutuksessa

Synteettinen puhe on tärkeä osa monia ihminen-robotti-vuorovaikutustilanteita [38] ja sillä on vaikutusta kuuntelijan asenteisiin ja käyttäytymiseen [39]. Monet robotin puheen piirteet vaikuttavat siihen, kuinka ihminen vastaanottaa puheen. Tällaisia piirteitä ovat esimerkiksi epäröinti, empiminen, sekä puhujan sukupuoli, kieli ja aksentti [36, 38, 39]. Epäröimätön robotti on karismaattisempi kuin epäröivä robotti [36], ja epäröiviin robotteihin verrattuna ihmiset suosivat epäröimättömiä robotteja ja seuraavat paremmin niiden antamia ohjeita [36]. Toisaalta synteettinen puhe, joka sisältää empimistä, tilkesanoja ja hajotettuja lausuntoja, on ihmismäisemmän, kohteliaamman ja älykkäämmän oloista, kuin synteettinen puhe, joka poistaa nämä epäsujuvuudet [35]. Näillä ominaisuuksilla ei kuitenkaan ole merkitystä puheen ymmärtämisen tai tehokkuuden kannalta [35]. Puhe, joka on samankaltaista kuin kuuntelijan oma puhe, on kuuntelijalle puoleensavetävämpää, kuin puhe, joka ei vastaa kuuntelijan puhetta lainkaan [39]. Lisäksi jos sanat, joita robotti käyttää ovat sanoja, joita kuuntelija käyttää usein omassa puheessaan, kuuntelija tuntee olonsa paremmaksi ja on halukkaampi vuorovaikutukseen robotin kanssa [40].

Puheen selkeys ja sujuvuus ovat välttämättömiä vuorovaikutuksessa, sillä jos kuuntelija ei ymmärrä tai saa selvää puhujasta, puhuttu viesti ei välity eteenpäin. Monimutkainen syntaksi, sekä liian yksinkertaiset sanat aiheuttavat väärinymmärtämisiä synteettisessä puheessa [41]. Foneettisesti erottuvat sanat selkeyttävät puhetta ja helpottavat sanojen erottelua [41], joten puhesynteesin tulisi käyttää ’helppoja’ sanoja. Puheen ymmärtämisen helpottamiseksi olisi myös syytä

käyttää kuuntelijan äidinkieltä, sillä ihminen ymmärtää äidinkieltään selvästi vierasta kieltä paremmin [41].

2.3.4. Toteutusten arviointi

Erilaisia toteutuksia voidaan arvioida yleensä kolmella eri tavalla: subjektiivisesti, objektiivisesti ja käyttäytymisen perusteella [42]. Subjektiivinen arviointi perustuu usein mielipidemittauksiin, objektiivinen mitattaviin ominaisuuksiin ja käytöksellinen arviointi sen arvioimiseen, suoriutuvatko koehenkilöt kyseisen systeemin avulla määrätystä tehtävästä paremmin tai huonommin kuin eri systeemin avulla [42].

On tärkeää ymmärtää, että ei ole olemassa universaalia standardia optimaaliselle synteettiselle puheelle [42], vaan on järkevämpää arvioida sitä, kuinka hyvin synteesi suoriutuu sovellusympäristössään ja vastaa sille asetettuihin odotuksiin. Kun puhesynteesi toteutetaan karikatyyriselle robotille, osan ihmisistä on todettu pitävän robottimaisemman kuuloista ääntä paremmin sopivana, kuin ihmismäistä ääntä [42]. Humanoidirobottien tapauksessa, yleinen synteettiselle puheelle asetettu vaatimus on, että puhe on humanoidin, mutta ei liian ihmismäisen äänen kaltaista [42]. Tällaiselle toteutukselle parhaat arviointimetodit ovat subjektiivisia ja käytöksellisiä, kuten havaitun kelvollisuuden arviointi tai tehtävästä suoriutumisen arviointi synteesin avulla [42].

Erilaisten robottisuiden arvioinnissa voidaan keskittyä moniin eri seikkoihin, kuten toteutusten realistisuuteen, eri asentovaihtoehtojen määrään ja liikkeiden sulavuuteen. Myös puheen ja suun liikkeen synkronisaation toteutusta voidaan arvioida osana suun arviointia. Puheen ja suun liikkeen synkronisaatiota arvioidaan usein objektiivisesti käyttämällä suurnopeuskameroita määrittämään suun asentojen ja äänteiden oikea synkronisaatio [7]. Toteutuksia voidaan myös arvioida subjektiivisin keinoin esimerkiksi mielipidemittausten avulla. On myös esitetty, että puheen ja suun liikeen synkronisaation parantamisen tulisi olla tärkeämpää, kuin jo olemassaolevan toteutuksen realistisuuden kasvattaminen [7].

2.3.5. Sosiaalisten robottien käytön etiikka

Robotiikka on kehittynyt siihen pisteeseen, että on erittäin mahdollista, että robotit tulevat toimimaan monissa sosiaalisissa rooleissa ihmisten elämissä [34]. Sosiaaliset robotit voivat toimia esimerkiksi opettajina, ohjaajina, hoitajina tai seuralaisina [34, 43]. Ihmiset tuntevat olonsa mukavammaksi sellaisen robotin seurassa, joka ei toimi ihmisten moraalikäsitystä vastaan, ja joka kunnioittaa ihmisten normeja ja arvoja [37]. On siis tärkeää miettiä sosiaalisten robottien käytön järkevyyttä, laillisuutta ja eettisyyttä eri aloilla ja eri kohderyhmien kannalta [34].

Lasten parissa toimivat robotit voivat olla esimerkiksi hoitajia, opettajia ja seuralaisia [34]. Suurimmat eettiset huolet lasten kanssa tekemisissä oleviin robotteihin liittyen ovat huoli lasten yksityisyydestä, kiintymyssuhteen muodostuminen robottiin, robotin harhaanjohtava olemus ja käytös sekä ihmiskontaktin menetys [34]. Jos lasten kanssa tekemisissä olevan robotin ’luonne’ on ystävällinen, voi lapsille syntyä illuusio siitä, että robotti on heidän ystävänsä ja aidosti välittäisi heistä [34]. Liiallinen

robottiystävän kanssa vietetty aika voi johtaa ihmiskontaktin vähenemiseen ja siten lapsen sosiaalisten taitojen puutteelliseen kehitykseen [34]. Lapsi voi esimerkiksi pahoinpidellä robottia eikä opi, että se olisi paha asia, jolla on seurauksia, koska robotti ei osaa erotella moraalista hyvää ja pahaa käytöstä toisistaan ja reagoida asianmukaisesti [34]. Suurimmat huolet yksityisyyden vaarantumisen lisäksi liittyvät siis negatiivisiin vaikutuksiin lasten sosiaalisissa taidoissa.

Vaikka aikuisten ihmisten voisi ajatella olevan tietoisia siitä, että robotit ovat vain koneita, totuus on se, että aikuisetkin voivat erehtyä uskomaan liikaa robotin kykyihin [34]. Jos robotti osaa matkia ihmisen käytöstä liian hyvin ja luoda illuusion tietoisuudesta ja myötätunnosta, se voi johtaa ihmisiä harhaan olemuksellaan, mikä voi puolestaan johtaa siihen, että ihmiset odottavat robotilta liikoja ja luottavat siihen enemmän kuin kannattaisi [8, 34]. Tämä voi johtaa siihen, että robotti asetetaan sille liian suuren vastuun alaiseksi, kuten opettajan rooliin, vaikka sillä ei ole samanlaisia valmiuksia toimia opettajana kuin ihmisellä [34].

Vanhusten parissa toimivat robotit voidaan jakaa yleisesti ottaen kolmeen ryhmään: arjessa avustaviin, terveyttä tarkkaileviin ja seuraa pitäviin robotteihin [43]. Suurimmat eettiset huolet näihin robotteihin liittyen ovat ihmiskontaktin väheneminen, tunne kontrollin menettämisestä ja esineellistämisestä, yksityisyyden menetys, henkilökohtaisen vapauden menetys, infantilisaatio sekä kysymys siitä, missä tilanteissa hoivattavat saavat itse kontrolloida robotteja [43]. Toisaalta etäläsnäolon mahdollistavien robottien käyttö voi olla tietyissä tilanteissa, kuten flunssakausien aikana järkevämpää kuin vierailevien ihmishoitajien käyttäminen. Näin terveysriskit sekä hoitajille, että hoidettaville pienenevät, kun fyysisiä kanssakäymisiä rajoitetaan.

Yleisellä tasolla kaikkia ihmisryhmiä yhdistävä huoli sosiaalisiin robotteihin liittyen on huoli yksityisyydestä. Kyseinen huoli johtuu siitä, että toimiakseen sosiaalisessa ympäristössä robotti kerää sensoreidensa avulla yksilöihin liittyvää henkiökohtaista dataa, joka voi joutua ulkopuolisten käsiin, jos sitä ei suojata ja säilytetä oikein [34, 43]. Monia ihmisiä myös huolettaa robottien käytön lainopillinen puoli, eli se, että voidaanko robotin tulkita olevan laillisessa vastuussa esimerkiksi sen hoidon alaisten ihmisten hyvinvoinnista ja hengestä [34].

3. TOTEUTUS

Tässä projektissa toteutettiin leuan liike ja puhesynteesi InMoov-robotille. InMoov-robotin päähän kuuluvat mm. kameroilla toimivat silmät, servolla liikutettava leuka, sekä äänen tunnistukseen ja tuottamiseen tarvittavat mikrofonit ja kaiuttimet. Toiminnallisuuden toteuttamiseen käytettiin hyväksi ROS-ympäristöä. Molemmista toiminnoista muodostettiin ROS-solmut, jotka kommunikoivat keskenään ROS:in julkaisija/tilaaja -mallin mukaisesti.

Leuan liike toteutettiin Dynamixel XL-320-servolla, jota kontrolloitiin Arduino UNO:lla. Servon ohjaaminen tapahtuu Arduinon välityksellä Dynamixelin 2.0-protokollaa noudattavien pakettien avulla. Puhesynteesinä toteutettiin konkatenaatiosynteesi ja formanttisynteesi. Konkatenaatiosynteesiä varten nauhoitettiin tarvittavat äänteet sekä mies- että naispuhujalla. Synteesin valittaviin asetuksiin kuuluvat puhuttu kieli, puhujan sukupuoli ja prosodia.

Näiden toteutettujen komponenttien avulla toteutettiin tervehtijärobotti, jonka tarkoitus on tervehtiä ihmisiä Oulun yliopiston Abipäivillä. Robotin vaadittaviin toiminnallisuuksiin kuuluvat ihmisten tervehtiminen sekä informaation jakaminen puheen avulla.

3.1. InMoov-robotti

InMoov (Kuva 6) on ensimmäinen avoimen lähdekoodin 3D-tulostettu ihmisen kokoinen humanoidirobotti. InMoov-robotin tulostettavat osat on suunnitellut ranskalainen Gael Langevin. InMoov-projekti aloitettiin vuonna 2012, jolloin myös ensimmäiset robotin osien 3D-mallit tulivat julkiseen käyttöön.

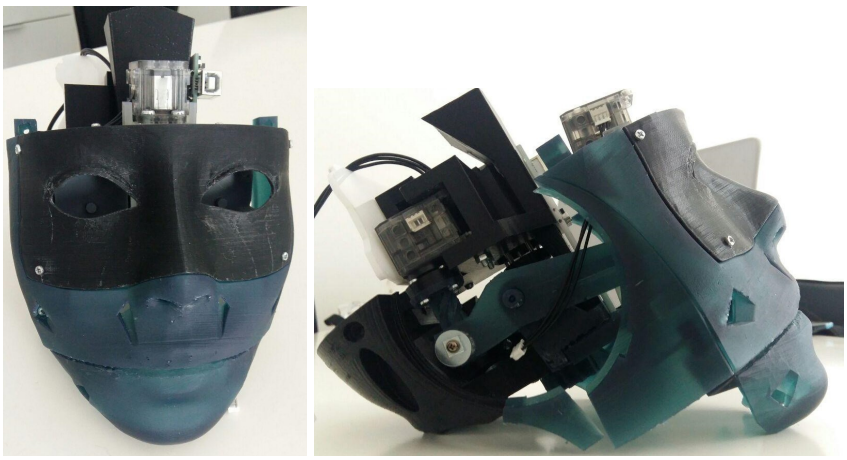
Kuva 6. InMoovin mallintaja Gael Langevin (vas.), sekä InMoov-robotti.[6]

[6]Kuva käyttäjältä Emmanuel Gilloz (https://www.flickr.com/people/watsdesign/), saatu lisenssin CC BY-SA 2.0 alaisena (https://creativecommons.org/licenses/by-nc/2.0/

Osat voidaan tulostaa 3D-tulostimilla joiden tilavuus on 12x12x12cm tai suurempi, ja sen tekijänoikeuslisenssi on Creative Commons Nimeä-EiKaupallinen (CC BY-NC), joten se on oivallinen kehitysalusta robottien tutkimiseen ja robotiikan oppimiseen.

Yleensä InMoov-robotille toteutetaan seuraavat ruumiinosat:

- kädet, joissa on liikutettavat nivelet sormissa, ranteissa, kyynär- ja olkapäissä;
- vartalo, joka yhdistää muut ruumiinosat keskenään, toimii joidenkin liike- ja kuva-antureiden sijaintina, ja pitää usein sisällään robotin laskennallisen laitteiston;
- pää (Kuva 7), jossa on kameroilla toimivat silmät, servolla liikutettava leuka, sekä äänen tunnistukseen ja tuottamiseen tarvittavat mikrofonit ja kaiuttimet;
- ja jalat, jotka ovat usein vain renkaat liikkumisen toteuttamiseen.



(a) InMoov edestä. (b) InMoov sivusta.

Kuva 7. Projektissa käytetyn InMoov robotin pää.[7]

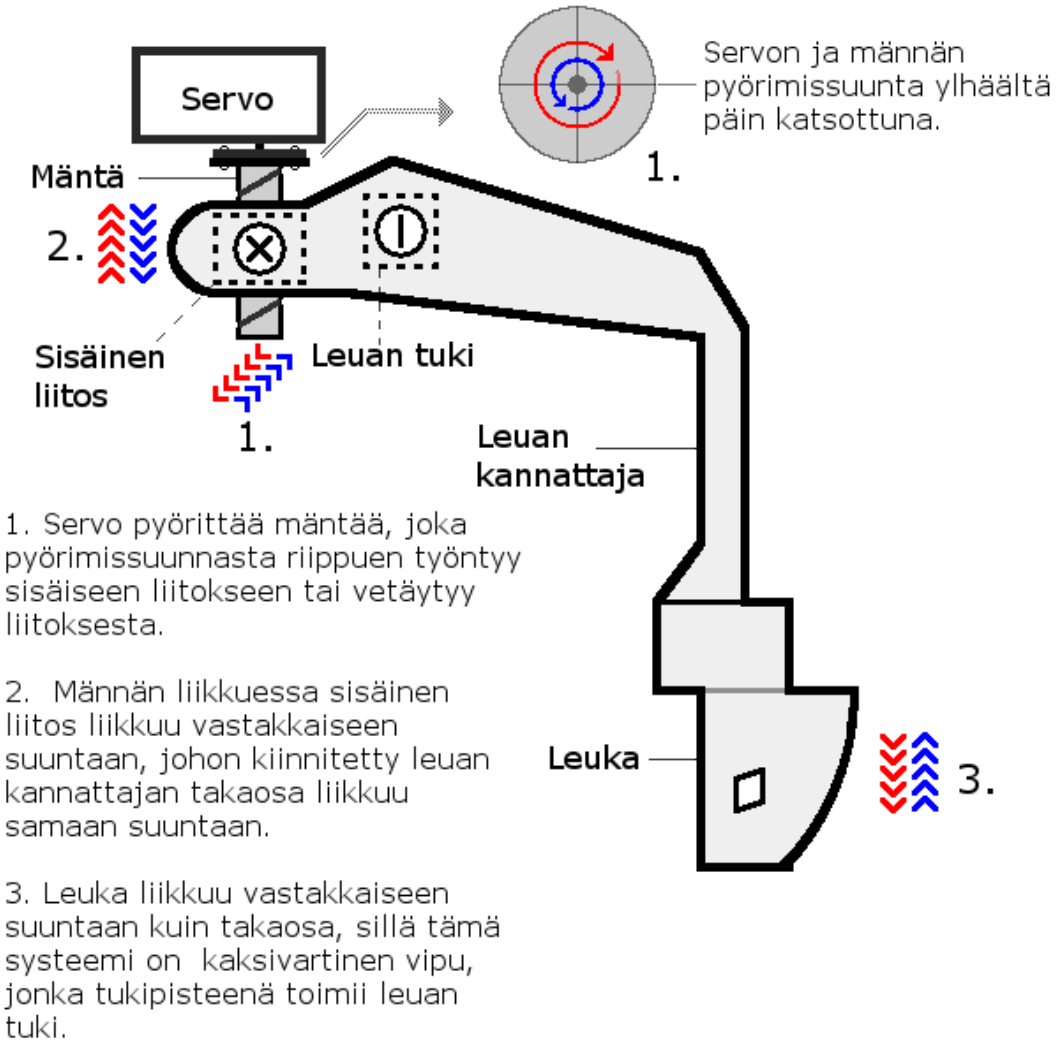
Vaikka jaloille löytyy omat 3D-tulostukset, niihin ei olla toteutettu toimivaa mekaniikkaa, joten kävely olisi nykyisillä toteutuksilla käytännössä mahdotonta.

InMoovin avoimen lähdekoodin ansiosta sen toteutuksia on monenlaisia, ja sen mahdolliset toiminnallisuudet ovat rajattomat, mutta tämän työn toteutuksessa vain pää, ja vielä tarkemmin leuan ja sen liikuttamisen toteutus ovat keskeisessä roolissa. Leuka koostuu seuraavista osista:

- yksi servo, jonka avulla leukaa liikutetaan;
- yksi mäntä, joka on kiinnitettynä (ruuveilla/liimalla) servoon;
- yksi sisäinen liitos, johon mäntä liitetään ja joka liikkuu pystysuunnassa männän pyörimisen mukaan;
- kaksi leuan tukea, jotka sijaisevat pään sivuilla ja tukevat leuan kannattajia;
- kaksi leuan kannattajaa, jotka kiinnitetään paikallaan pysyviin tukiin, liitoksen molemmille puolille, ja leukaan;
- sekä yksi leuka.

Kuva 8 havainnollistaa ja selittää, kuinka leuan liikkuminen toteutuu servon avulla.

[7]Kuvat ottanut Tiia Leinonen. Kuville asetettu lisenssi CC BY 2.0 (`https://creativecommons.org/licenses/by/2.0/`).



Kuva 8. Leuan liikuttaminen servolla.[8]

3.2. ROS-ympäristö

ROS (Robot Operating System) on avoimen lähdekoodin metakäyttöjärjestelmä, jota robotti käyttää. Se tarjoaa palveluita, jota käyttöjärjestelmältä voi odottaa, kuten matalan tason laitteistokontrollin, usein käytetyn toiminnallisuuden implemennoinnin, prosessien välisen viestien välityksen ja pakettien käsittelyn.

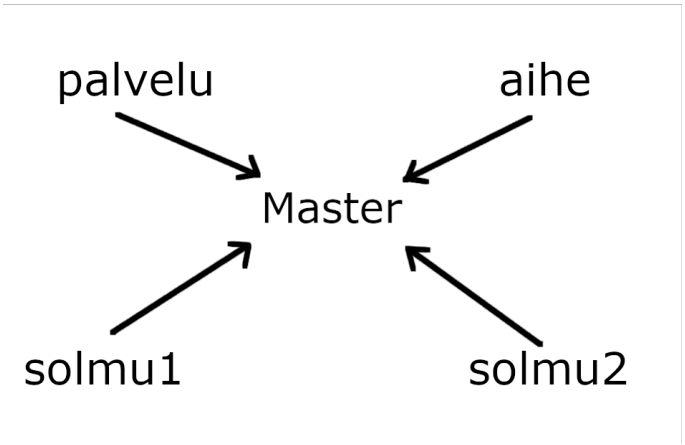
Paketit ovat pääkeino ohjelmistokomponenttien järjestämiseen ROS:issa. Paketti voi sisältää ROS-prosesseja, kirjastoja, dataa, konfiguraatiotiedostoja tai mitä tahansa tiedostoja, jotka muodostavat yhdessä toimivan kokonaisuuden.

ROS:in toiminnallisuuden muodostavat ROS-prosessit. Yhdessä dataa työstävät prosessit muodostavat yhdessä ROS:in laskentagraafin (computation graph), joka on käytännössä prosessien vertaisverkko. Siihen liittyviä käsitteitä:

[8]Kuvan tehnyt Santtu Käpylä. Kuvalle asetettu lisenssi CC BY 2.0 (`https://creativecommons.org/licenses/by/2.0/`).

- Node eli solmu on prosessi, joka suorittaa laskennallisuutta. Robottia ohjaava systeemi koostuu usein monesta solmusta, joista jokaisella on oma tehtävänsä. ROS-solmut toteutetaan ROS:in asiakaskirjastoilla, kuten roscpp:llä tai rospy:llä.
- Master on tiedosto, joka tarjoaa nimien rekisteröinnin ja haun. Ilman sitä solmut eivät kykenisi löytämään toisiaan tai vaihtamaan viestejä tai kutsumaan palveluja.
- Viestit ovat keino, jolla ROS-solmut kommunikoivat keskenään. Ne koostuvat dataa sisältävistä kirjoitetuista kentistä.
- Topic eli aihe on nimi, jolla tunnistetaan viestien sisältö. Viestit ohjataan oikeille solmuille julkaisija/tilaaja -järjestelyllä. Solmu, joka on kiinnostunut tietystä datasta alkaa tilaajaksi asianmukaiselle aiheelle, jolloin kun aihekanavalle julkaistaan dataa, tilaajasolmu voi sen vastaanottaa. Samalle aihekanavalle voi olla yhtäaikaisesti monta tilaajaa ja julkaisijaa, ja yksi solmu voi toimia julkaisijana usealle aihekanavalle.
- Service eli palvelu on vaihtoehto julkaisija/tilaaja -mallille. Joskus tarvitaan kahden solmun välistä pyyntö/vastaus -vuorovaikutusta, johon julkaisija/tilaaja -malli ei kykene yksisuuntaisen tiedonvälityksensä vuoksi. Pyyntö ja vastaus määritellään pariksi viestirakenteita: yksi pyynnölle ja toinen vastaukselle. Palvelun tuottava solmu tarjoaa palvelua ja asiakas käyttää palvelua lähettämällä palvelijasolmulle palvelupyynnön viestillä ja odottamalla vastausta.

Yhdessä edellämainitut konseptit toimivat kuvan 9 esittämällä tavalla: Master toimii laskentagraafin nimipalvelimena ja rekisteröi aiheiden ja palveluiden tiedot solmuja varten. Solmut kommunikoivat Master:in kanssa ja raportoivat omat tietonsa sille. Näin solmut voivat tarvittaessa löytää toisensa kommunikoimalla Masterin kanssa.

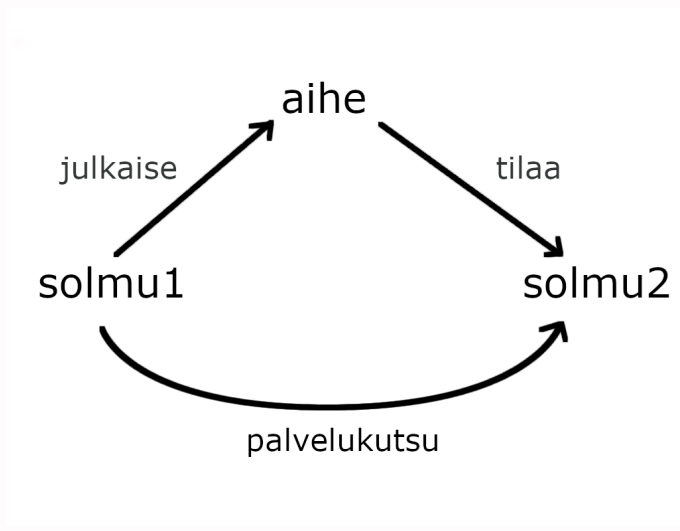


Kuva 9. Master:in toiminta visualisoituna. Master toimii nimipalvelimena solmujen, aiheiden ja palvelujen informaatiolle. Nuolet kuvaavat tiedon välittymistä Masterille.[9]

Solmut kommunikoivat keskenään suoraan kuvan 10 havainnollistamalla tavalla löydettyään toisensa Masterin avulla. Solmut, jotka tilaavat aiheita pyytävät yhteyden muodostamista julkaisijasolmuilta, ja muodostavat kyseisen yhteyden

[9]Kuvan tehnyt Tiia Leinonen. Kuvalle asetettu lisenssi CC BY 2.0 (`https://creativecommons.org/licenses/by/2.0/`).

ennalta sovitulla yhteysprotokollalla, joka ROS:issa on yleisimmin TCPROS, joka käyttää standardinmukaisia TCP/IP-pistokkeita.



Kuva 10. Solmujen kommunikointi keskenään tapahtuu joko julkaisija/tilaaja -mallilla tai palvelukutsuilla. Tässä kuvassa nuolet kuvaavat viestejä.[10]

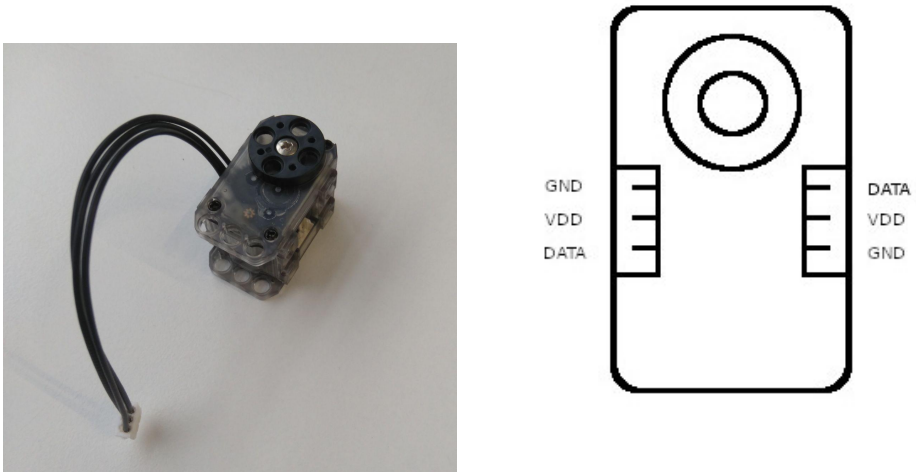
3.3. Käytetyt komponentit

Leuan liikkeen tuottamiseen käytettiin Dynamixel XL-320 -servoa. Servoa ohjataan muuttamalla sen rekisteriarvoja lähettämällä sille käskypaketteja ja vastaanottamalla siltä tilapaketteja, jotka noudattavat Dynamixelin 2.0-protokollaa. Käskypaketit (Taulukko 3) sisältävät otsikon (HEADER), servon tunnisteen (ID), viestin pituuden (LEN), toimintakäskyn (INS), parametrit käskyä varten (PARAM) sekä tarkistussumman (CRC), jolla voidaan varmistaa paketin eheys. Tilapaketeissa (Taulukko 4) on mukana lisäksi kenttä virhetietoja varten (ERR).

Luvut ovat paketissa heksadesimaalimuodossa, ja kentän ollessa suurempi kuin yksi tavu, se merkitään little endian -muodossa eli vähiten merkitsevä tavu ensin, kuten esimerkiksi kentän CRC tapauksessa: ensin tulee CRC_L, eli vähiten merkitsevä tavu, ja sitten CRC_H, eli eniten merkitsevä tavu.

Servon XL-320 tapauksessa pakettien otsikko on aina muotoa [0xFF,0xFF,0xFD,0x00] ja tilapakettien käsky on aina 0x55. Yhden servon tapauksessa tunnisteena voidaan käyttää servon vakiotunnistetta 0x01 tai broadcast-tunnistetta 0xFE. Useamman servon tapauksessa servoille täytyy antaa toisistaan eroavat tunnisteet, jotta tietyt käskyt välittyvät oikeille servoille. Paketin parametrien lukumäärä ja parametrien arvot riippuvat toimintakäskystä, joka voi olla esimerkiksi 0x03, eli kirjoitus. Viestin pituus saadaan lisäämällä parametrien lukumäärään luku kolme.

[10]Kuvan tehnyt Tiia Leinonen. Kuvalle asetettu lisenssi CC BY 2.0 (`https://creativecommons.org/licenses/by/2.0/`).



(a) (b)

Kuva 11. Dynamixel XL-320 -servo.[11] Kuvassa (a) servo käytännössä ja kuvassa (b) servon pinnien esitys.

Taulukko 3. Käskypaketin muoto. Luvut heksadesimaaleina.

HEADER	ID	LEN	INS	PARAM	CRC
[FF,FF,FD,00]	ID	[L_L, L_H]	INS	[P1,..., PN]	[CRC_L, CRC_H]

Taulukko 4. Tilapaketin muoto. Luvut heksadesimaaleina.

HEADER	ID	LEN	INS	ERR	PARAM	CRC
[FF,FF,FD,00]	ID	[L_L, L_H]	55	E	[P1,..., PN]	[CRC_L, CRC_H]

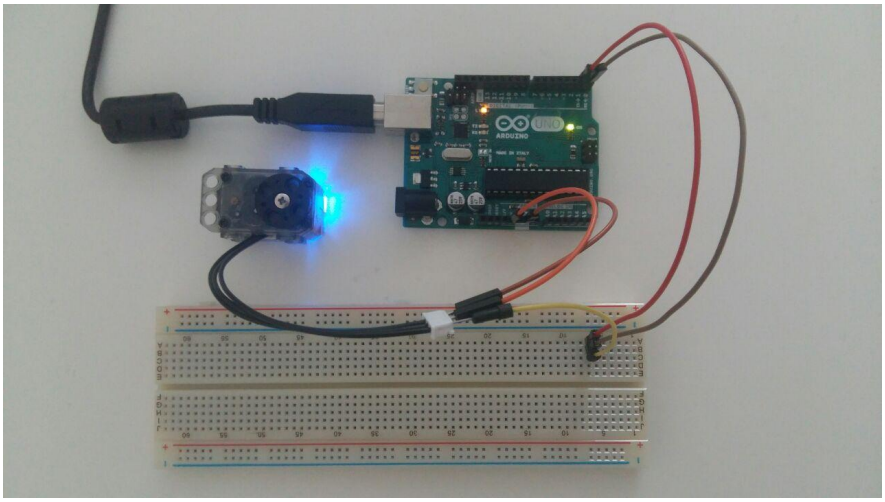
Servolla on vain yksi datalinja, jota pitkin kulkevat sekä data servolle, että servolta takaisin, eli se on ns. epäsynkroninen vuorosuuntainen datalinja. Erisuuntaisten datapakettien välillä on oletuksena 500 mikrosekunnin viive. Datalinjalla kulkeva data voidaan esittää notaatiolla 8N1, mikä tarkoittaa kahdeksaa databittiä, yhtä aloitusbittiä, yhtä lopetusbittiä ja ei yhtään pariteettibittiä.

Servon muisti koostuu EEPROM- ja RAM-osioista. RAM-muistilla oleva data palautuu oletusarvoihinsa servon virran katketessa, mutta EEPROM-muistilla oleva data säilyttää arvonsa. Tehdasasetusten mukaisesti servon datalinjan nopeuden vakioarvo on 1000000 baudia, ja servon tunniste on 1, mutta niitä voi muuttaa tarvittaessa.

Servon kontrolloimiseen käytettiin Arduino UNO:a. Kuva 12 esittää testeissä käytetyn kytkennän ja kuva 11 servon pinnien järjestyksen. Servon pinni DATA kytkettiin koekytkentälevyn kautta Arduinon pinneihin 0 ja 1, pinni VDD Arduinon pinniin 5V ja pinni GND Arduinon pinniin GND. Lopullisessa ROS-komponentissa

[11]Kuvat ottanut ja tehnyt Tiia Leinonen. Kuville asetettu lisenssi CC BY 2.0 (`https://creativecommons.org/licenses/by/2.0/`).

kytkentä yksinkertaistui niin, että servon pinni DATA voitiin kytkeä suoraan Arduinon pinniin 1.



Kuva 12. Testeissä käytetty servon ja Arduinon välinen kytkentä.[12]

3.4. Omat ohjelmistokomponentit

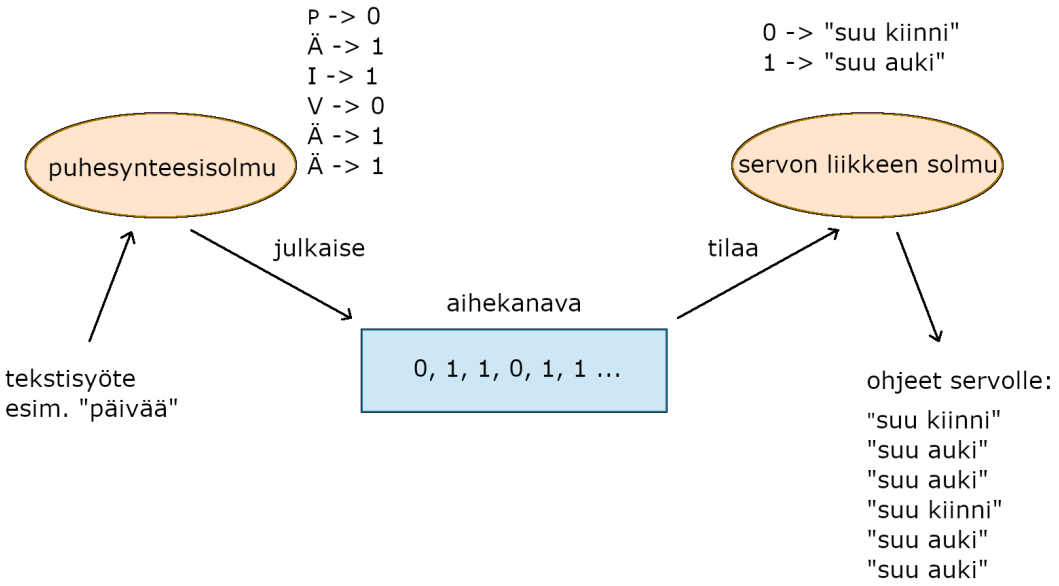
Työssä toteutettiin itse kaksi ROS-komponenttia, joista toinen on leuan liikkeelle ja toinen puhesynteesille. Komponentit kommunikoivat keskenään kuvan 13 esittämällä tavalla ROS:in julkaisija/tilaaja -mallia käyttäen.

3.4.1. Puhesynteesi

Robotille toteutettiin sekä TTS-konkatenaatio-, että TTS-formanttisynteesi, jotta puhesynteesiä voidaan vaihdella tilanteen mukaan sopivaksi. Konkatenaatiosynteesi valittiin, sillä se säilyttää puhutun äänen luonnollisuuden, formanttisynteesi siksi, että sillä voidaan tuottaa sulavaa puhetta, ja tekstistä puheeksi -muunnosta käyttämällä robotilla on mahdollisuus puhua ’mitä tahansa’ sille annetaan syötteeksi. Koodi on kirjoitettu Python-ohjelmointikielellä, jonka versio on 2.7.17, sillä 2.7.17 on viimeisin ROS:in rospy-kirjastoa tukeva Python-versio.

Konkatenaatiosynteesi tarvitsee toimiakseen nauhoitettuja äänipätkiä, joista puhe muodostetaan. Nauhoitus toteutettiin suomenkielisille äänteille lausumalla radioaakkoset ja leikkaamalla niistä sanassa esiintyvä äänne. Lisäksi nauhoitettiin äänet myös sanaväleille, lopetusmerkeille, suomen kielen diftongeille ja pitkitetyille monoftongeille. Sanaväleissä ja lopetusmerkeissä käytetään hiljaisuutta, jotka kestävät 0,15s ja 0,25s, tässä järjestyksessä. Kaikki äänitykset toteutettiin sekä nais-, että miespuhujalla.

[12]Kuvan ottanut Tiia Leinonen. Kuvalle asetettu lisenssi CC BY 2.0 (`https://creativecommons.org/licenses/by/2.0/`).



Kuva 13. Solmujen välinen kommunikaatio pelkistettynä.[13]

Formanttisynteesiä varten nauhoitetuista lauseista katsottiin jokaisen äänteen neljä ensimmäistä formanttia, monoftongeja ja diftongeja lukuunottamatta. Lisäksi katsottiin äänteiden kestot ja formanttien kaistanleveydet. Hiljaisissa merkeissä käytettiin samaa pituutta kuin konkatenaatiosynteesissä ja niiden formanteiksi ja kaistanleveyksiksi asetettiin nollat. Lisäksi valittiin puheelle perustaajuudeksi (F0) 105 Hz.

Puhesynteesin ja ROS:in yhteensopivuus toteutettiin kahdella julkaisijasolmulla ja kuudella tilaajasolmulla. Julkaisijasolmut käyttävät aihekanavavia ’tts_ready’ ja ’move_servo’ ja tilaajasolmut käyttävät aihekanavia ’text’, ’language’, ’gender’, ’prosody’, tts_type, ja ’servo_ready’.

Toteutetulla TTS:llä on vaihdettavia asetuksia, joita muuttamalla voidaan vaikuttaa puheen muodostamiseen. Asetuksiin kuuluvat kieli, sukupuoli, prosodia, sekä käytetty puhesynteesi. Asetuksien vaihto tapahtuu ROS-tilaajasolmuja ’language’, ’gender’, ’prosody’, ja ’tts_type’ hyödyntäen, ottamalla merkkijono-muuttujan (String) haluttuun aihekanavaan ja asettamalla tämän suoraan asetukseksi. Käytetyn puhesynteesin valinnalla voidaan vaikuttaa siihen minkä tyyppistä synteesiä käytetään. Valittu kieli ja sukupuoli vaikuttavat siihen, mitä nauhoitettuja ääniä käytetään puhesynteesin tuottamisessa ja kieli vaikuttaa lisäksi myös siihen kuinka teksti tulkitaan. Suomenkielisessä TTS:ssä funktio ottaa tekstistä suoraan äänteet kirjaimista, mutta tämä ei ole mahdollista kaikissa kielissä kuten englannissa, jossa lausuminen riippuu suuresti käytetystä sanasta. Prosodian asetuksilla voidaan vaikuttaa tuotetun puheen prosodiaan. Esimerkkejä mahdollisista prosodisista asetuksista ovat mm. äänen voimakkuus, tempo, sekä kirjainten ja sanojen painotus. Puhesynteesin asetuksien vaihdolla ei ole vaikutusta, jos tekstiä ollaan jo muuttamassa puheeksi, sillä asetukset

[13]Kuvan tehnyt Tiia Leinonen. Kuvalle asetettu lisenssi CC BY 2.0 (`https://creativecommons.org/licenses/by/2.0/`).

katsotaan heti TTS:n alussa, mutta muutokset otetaan huomioon seuraavassa TTS-pyynnössä.

Toteutettu TTS-funktio ottaa syötteekseen tekstin merkkijono-tyyppisenä aihekanavasta 'text'. Syötteen saatuaan funktio tarkistaa onko aikaisempi puhesynteesi kesken ja jos näin on, uutta synteesiä ei toteuteta, jos ei, jatketaan normaalisti. Tämän jälkeen valitaan käytettävä puhesynteesi asetusten mukaisesti. Oletuksena toimii konkatenaatiosynteesi, sillä sillä tuotettu puhe on luonnollisemman kuuloista kuin formanttisynteesin tuottama puhe. Formanttisynteesiä voitaisiin käyttää esimerkiksi tilanteessa, jossa systeemin muisti on rajallinen, sillä formanttien säilöminen vaatii vähemmän muistia kuin nauhoitettujen äänien.

Funktio 1. Ääniaallon muodostus formanteista

Syötteet: kirjain **k**, kirjaimen tyyppi **t**, ja näytteenottotaajuus **Fs**
Lähtö : Ääniaalto **puhe**

$Formantit \leftarrow numpy.array(formantit[\mathbf{t}][\mathbf{k}][0:4])$
$Bandwidth \leftarrow [20, 30, 50, 60]$
$kesto \leftarrow numpy.array(formantit[\mathbf{t}][\mathbf{k}][4])$
$samples \leftarrow math.floor(kesto * \mathbf{Fs})$
$radius \leftarrow numpy.exp(-numpy.pi * Bandwidth/\mathbf{Fs})$
$\theta \leftarrow (2 * numpy.pi * Formantit)/\mathbf{Fs}$
$navat \leftarrow radius * numpy.exp(j * \theta)$
$[B, A] \leftarrow scipy.signal.zpk2tf(0, navat, 1)$
$F0 \leftarrow 105$
$w0T \leftarrow 2 * numpy.pi * F0/\mathbf{Fs}$
$nharm \leftarrow math.floor((\mathbf{Fs}/2)/f0)$
$sig \leftarrow numpy.zeros(samples)$
$n \leftarrow numpy.arange(samples)$
if $Formantit[0] \neq 0$ **then**
    for $i \leftarrow 0$ **to** $nharm$ **do**
        $sig \leftarrow sig + scipy.signal.sawtooth(i * w0T * n)$
    end
end
$sig \leftarrow sig/numpy.max(sig)$
puhe $\leftarrow scipy.signal.lfilter(numpy.array([1,0]), A, sig)$

Formanttisynteesi muodostaa ääniaallon jokaiselle äänteelle erikseen ja yhdistää ääniallot muodostaen puheäänen syötetystä tekstistä. Näytteenottotaajuus (Fs) on asetettu arvoon 8192 Hz. Ääniaallon muodostaminen yhdelle kirjaimelle nähdään funktiosta 1. Funktiossa käytetetty formantit-sanakirja on jakautunut 'vokaali'- ja 'konsonantti'-sanakirjoihin, jotka sisältävät kaikki äänteet, ja konsonantteihin on lisätty myös hiljaiset merkit, kuten ','. Sanakirjamuuttujina toimivat listat, joissa on kunkin äänteen formantit, sekä äänteen kesto, muodossa [F1, F2, F3, F4, t]. Näitä arvoja käyttäen lasketaan navat ja näytteiden määrä. Tämän jälkeen muodostetaan signaali F0:n mukaan, joka on muodoltaan kolmioaalto. Kolmioaalto valittiin, sillä siinä ei häviä informaatiota, ja näin ollen ääni on selkeää, toisin kuin kantti-, sini- ja kosiniaalloissa. Hiljaisissa äänteissä tätä ei tehdä, joten

arvot pysyvät nollassa, mikä vastaa tyhjää ääntä. Signaali suodatetaan formanteista saatujen nimittäjän polynomisten kertoimien mukaan. Kun äänteen ääniaalto on tehty, se lisätään muiden jo tehtyjen aaltojen perään. Kun kaikki äänteet on käyty läpi, aallon amplitudia pienennetään kymmenesosaan alkuperäisestä, äänen särinän pienentämiseksi. Funktiossa käytetään numpy-kirjastoa[14] taulukoiden tekemiseen, scipy-kirjastoa[15] signaalin napojen analysointiin ja puhesignaalin tuottamiseen, sekä math-kirjastoa[16] arvojen pyöristämiseen.

Konkatenaatiosynteesi valitsee käytettävät äänipätkät valittujen asetusten mukaisesti, tai oletusarvoisesti suomenkielisen miehen äänipätkät, jos asetuksia ei olla asetettu tai jos valittuja asetuksia ei olla toteutettu. Tämän jälkeen alkaa äänitiedoston luominen AudioSegment-kirjaston avulla. Valitusta kielestä ja prosodiasta riippuen, konkatenaatiosynteesi yhdistää äänteitä tekstin mukaan. Kun teksti on käyty läpi, yhdistetyistä äänistä muodostettu tiedosto tallennetaan myöhempää käsittelyä varten ja funktio lähettää aihekanavalle 'tts_ready' loogisesti toden totuusarvomuuttujan (Boolean) .

Puheen tuottaminen toteutettiin erillisenä funktiona, joka ottaa pääohjelmalta syötteekseen totuusarvomuuttujan aihekanavalta 'servo_ready'. Muuttujan ollessa tosi, ja jos puhesynteesi on käynnissä, funktio jatkaa toimintaansa, muulloin mitään ei tapahdu. Konkatenaatiosynteesissä funktio avaa muodostetun WAV-tiedoston SoundFile-kirjastoa käyttäen, minkä jälkeen lasketaan tiedoston ajallinen kesto, joka jaetaan TTS-funktion syötteksi tulleen merkkijonon pituudella, minkä avulla saadaan tahti. Tämän jälkeen funktio tuottaa äänitiedoston äänen pygame-kirjaston funktion pygame.mixer.music.play(1)[17] avulla. Formanttisynteesissä ääni muodostetaan pygame.sndarray-kirjaston[18] avulla ja muodostettu ääniaalto soitetaan 16-bittisenä näytteistystaajuuden mukaan. Samanaikaisesti funktio lähettää servolle tahdin mukaan käskyjä liikuttaa suuta käyttäen aihekanavaa 'move_servo', riippuen siitä mikä kirjain on menossa TTS:lle syötetyssä merkkijonossa. Kun kaikki merkit on käyty, merkitään puhesynteesi käydyksi, ja funktio päättyy. Tuotettu ääni ja servolle lähetetyt käskyt alkavat, toimivat, ja päättyvät samanaikaisesti.

3.4.2. Leuan liike

Robotin leuan liikkeen kontrolloimiseksi toteutettiin ROS-tilaajasolmu, joka asetetaan kuuntelemaan aihekanavaa, jolle puhesynteesistä vastaava solmu julkisee parametreja, jotka kertovat milloin servon on liikuttava ja suun avauduttava.

Servon toiminnan testaus aloitettiin lähettämällä servolle Dynamixelin 2.0-protokollan mukaisia käskypaketteja servon sisäisen LED:in vilkuttamiseksi sekä itse servon liikuttamiseksi haluttuihin asentoihin. Pakettien tarkistussummien laskemiseen käytettiin Robotis:in tarjoamaa update_crc -funktiota[19]. Kun toimintaa oli testattu tarpeeksi, aloitettiin tilapakettien lukuoperaation työstäminen.

[14] https://numpy.org/
[15] https://scipy.org/
[16] https://docs.python.org/3/library/math.html
[17] https://www.pygame.org/docs/ref/mixer.html
[18] https://www.pygame.org/docs/ref/sndarray.html
[19] http://emanual.robotis.com/docs/en/dxl/crc/

Lukuoperaatiota varten toteutetaan perinteisesti laitteistotason piiri, jolla servon data muunnetaan yksilinjaisesta vuorosuuntaisesta kaksisuuntaiseksi, mutta kyseistä tehtävää varten löydettiin valmis Arduino-kirjasto, XL320[20], joka hoitaa muunnoksen ohjelmistotasolla. Kyseistä kirjastoa käyttämällä voitiin lukea robotin päähän kiinnitetyn servon asento ja selvittää kokeilemalla suun mahdolliset minimi- ja maksimiasennot, sekä liikuttaa servoa haluttuun asentoon yksinkertaisemmin. Kuva 12 esittää testeissä käytetyn kytkennän ja taulukko 5 toteutuksessa käytettävät servon asennot sekä niihin liittyvät tarkistussummat.

Taulukko 5. Servon asennot ja niihin liittyvät tarkistussummat.

Suun asento	Luettu arvo	CRC_L	CRC_H
kiinni	900 = 0x0384	0x42	0xC5
auki	990 = 0x03DE	0x59	0x8D

Kun toteutuksesta alettiin tekemään ROS-komponenttia, huomattiin kuitenkin yhteensopivuusongelma kirjaston XL320 sekä ROS:in omien otsikkotiedostojen kanssa, jolloin päätettiin jättää lukuoperaatio toteutuksesta pois ja palata aikaisempaan versioon, eli suoraan pakettien lähettelyyn. Tämä oli mahdollista, koska päähän kiinnitetyn servon ääriasennot oli jo saatu luettua, ja koska servo on fyysisesti kiinni päässä, se ei pääse liikkumaan vapaasti, jolloin luetut ääriasennot pysyvät samoina ja niitä on turvallista käyttää. Koodin keventämiseksi ja Arduinon muistin säästämiseksi myös funktio update_crc jätettiin pois, sillä tarvittavia käskypaketteja on lopullisessa toteutuksessa vain kaksi erilaista ja niiden tarkistussummat pysyvät samoina. Servon asennon lukemista ja tarkistussummien laskemista varten ROS-pakettiin sisällytettiin kuitenkin tulevaisuuden varalta niihin tarvittavat Arduino-ohjelmat.

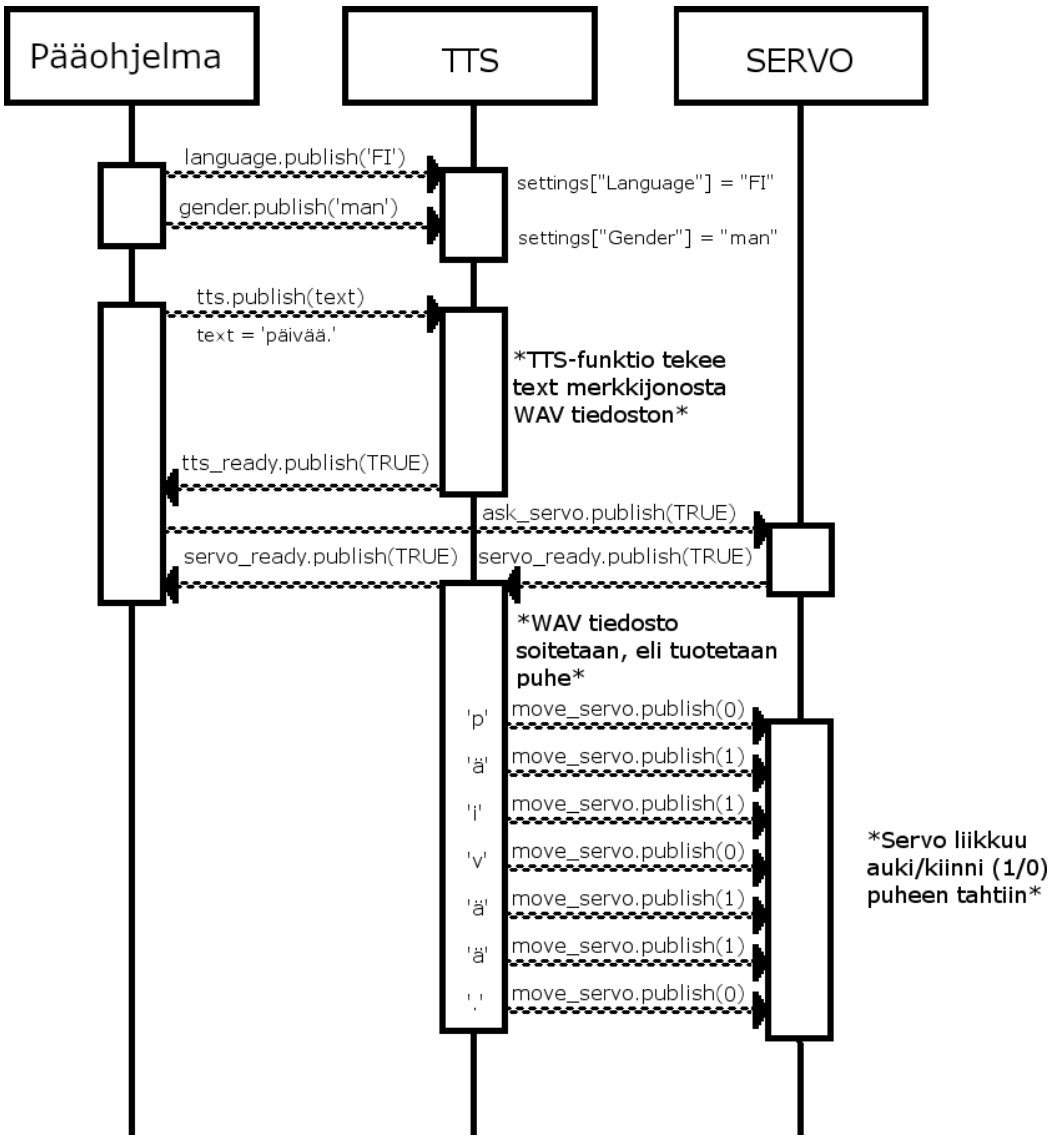
Servon toimintaa varten tehtiin lopuksi ROS-tilaajasolmu, joka tilaa aihekanavan, jolle puhesynteesisolmu julkaisee lukuja 0 ja 1. Luku 0 tarkoittaa 'suu kiinni' ja luku 1 tarkoittaa 'suu auki'. ROS-solmu saa parametrikseen jommankumman luvuista, jonka avulla sen sisäinen funktio servo_cb() valitsee oikean paketin lähetettäväksi servolle. Servon mahdollisia asentoja oli aluksi kolme, mutta leuan aukeamiskulma on niin pieni, että keskimmäinen asento, joka avasi suun raolleen ei näyttänyt testeissä hyvältä, jolloin se jätettiin pois.

3.5. Sovellusympäristö

Hyödyntäen edellämainittuja ohjelmisto- ja ROS-komponentteja voidaan tuottaa tervehtijärobotti, joka kykenee tervehtimään ihmisiä ja kertomaan tarvittavaa informaatiota. Käytännössä robotti siis tervehtii ihmisiä sanoen esimerkiksi: 'Tervetuloa Oulun yliopiston abipäiville! Ottakaa tästä tietotekniikan opetussuunnan esite. Kiitos näkemiin!'

Kuvan 14 tilanne esittää yksinkertaisen tervehdysrobotin toiminnan. Ensin valitaan asetukset kielelle ja sukupuolelle käyttäen ROS:in tilaaja- ja julkaisusolmuja pääohjelmalta TTS-funktiolle, joka vaihtaa nämä suomen kieleksi ja miehen ääneksi.

[20]XL320- kirjaston Github: `https://github.com/hackerspace-adelaide/XL320`



Kuva 14. Sovelluksen toiminta kaaviolla.[21]

Tämän jälkeen pääohjelma lähettää TTS-funktiolle käskyn tehdä WAV-tiedosto merkkijonosta ’Päivää.’. Tiedoston valmistuttua TTS-funktio lähettää pääohjelmalle varmistuksen tiedoston luonnista. Sen jälkeen puhesynteesi käynnistyy ja TTS-funktio lähettää servolle puheen tahtiin käskyjä liikuttaa suuta asianmukaisesti, kunnes kaikki äänteet on sanottu.

Tätä sovellusta voidaan myös hyödyntää osana laajempaa kokonaisuutta. Toiminnallisuuteen voidaan esimerkiksi liittää komponentti, joka tunnistaa ja seuraa ohikulkevia ihmisiä ja kääntää robotin katseen puhuteltavaa henkilöä kohti, jolloin vuorovaikutustilanteesta tulee luonnollisempi.

[21]Kuvan tehnyt Santtu Käpylä. Kuvalle asetettu lisenssi CC BY 2.0 (`https://creativecommons.org/licenses/by/2.0/`)

3.6. Tulosten arviointi ja vertailu

Toteutuksen arviointi jaettiin kahteen osa-alueeseen: puhesynteesin laadun, sekä puheen ja leuan liikkeen synkronisaation arviointiin.

Arviointitapa sekä synteesin laadun, että synkronisaation suhteen oli subjektiivinen, ja arviointeja varten kehitettin omat arviointiasteikot, joiden perusteella osatoteutuksille voitiin antaa arvosanat. Lopullinen arvosana konkatenaatiosynteesille oli 3/5, formanttisynteesille 2/5 ja puheen ja leuan liikkeen synkronisaatiolle 3/4.

3.6.1. Puhesynteesin arviointi

Koska InMoov on humanoidirobotti, sen puheen on oltava humanoidin, mutta ei liian ihmismäisen kuuloista, kuten kappaleessa 2.3.4 on esitetty. Tässä projektissa puhesynteesille toteutettu arviointi oli subjektiivista, ja puheen arviointi tapahtui asteikolla yhdestä viiteen, joille määritettiin kriteerit, jotka synteesin on täytettävä yltääkseen tiettyyn arvosanaan. Taulukko 6 esittää arvioinnissa käytetyn asteikon.

Taulukko 6. Puhesynteesin arviointiin käytettävä asteikko.

Arvosana	Vaatimukset arvosanalle
1	Puheen välittämä viesti ei ole ymmärrettävissä lainkaan. Useita äänteitä ei erota toisistaan tai ne puuttuvat kokonaan. Puheessa esiintyy erittäin paljon häiritsevää äänen hyppelyä.
2	Puheen välittämä viesti jää epäselväksi. Äänteissä on puutteita ja ne menevät keskenään sekaisin. Puheessa esiintyy paljon häiritsevää äänen hyppelyä.
3	Puheen välittämä viesti on pääpiirteittäin ymmärrettävissä. Vain tietyissä äänteissä on puutteita. Puheessa esiintyy jonkin verran häiritsevää äänen hyppelyä.
4	Puheen välittämä viesti on ymmärrettävissä. Lähes kaikki äänteet erottuvat selvästi. Puheessa esiintyy vain hieman häiritsevää äänen hyppelyä.
5	Puhe on selkeää ja välitetty viesti helposti ymärrettävissä. Kaikki äänteet erottuvat selvästi. Puheessa ei esiinny häiritsevää äänen hyppelyä.

Arvioinnissa käytettiin hyväksi neljää eri testilausetta, jotka sisälsivät monipuolisesti eri merkkejä kattavaa testausta varten. Molemmat synteesit ja eri puhujavaihtoehdot testattiin kaikilla lauseilla. Taulukko 7 esittää testaamiseen käytetyt lauseet. Testilauseet olivat tarkoituksella haastavia, vaikka puhesynteeseissä suositellaankin käytettäväksi ’helppoja’ sanoja.

Ensimmäinen versio konkatenaatiosynteesistä sai arvosanan kaksi, sillä puhe oli epäselvää ja äänteissä paljon puutteita. Arvion perusteella nauhoitettiin ja leikattiin uudet äänteet toista versiota varten, ja kyseisellä toteutuksella päädyttiin molemmilla

Taulukko 7. Puhesynteesin testaamiseen käytetyt lauseet.

Nro.	Lause
1	Kuningas Yrjö seitsemäs matkusti Islantiin ristiretkellään.
2	Bertta keräilee xylofoneja, ja Gabriel Zorro-figuureja.
3	Äiti ui joessa hyisenä syysyönä.
4	Sipsi ja dippi ovat leffaeväitä.

äänillä arvosanaan kolme. Miesäänellä äänteet /k/, /t/ ja /p/ jäivät puutteelliseksi ja naisäänellä äänteet /t/ ja /p/. Naisäänellä myös osa vokaaleista sisälsi äänenkorkeuden hyppelyä.

Formanttisynteesillä saavutettiin arvosana kaksi. Puheen välittämä viesti jäi epäselväksi, mutta vokaalien erottaminen onnistui puheesta helposti. Konsonantit jäivät suurelta osin epäselväksi.

3.6.2. Leuan liikkeen synkronisaation arviointi

InMoov-robotin leuan toteutus on monimutkaisuudeltaan verrattavissa FLASH:in suun toteutukseen, jossa suu kykenee vain pystysuuntaiseen liikkeeseen. Koska InMoov on kuitenkin humanoidirobotti, näin rajoittunut suun liike on sille liian yksinkertainen toteutus, mikä vaikeutti hyvän synkronisaation toteuttamista.

Puhesynteesin ja leuan liikkeen synkronisaatiolle toteutettu arviointi oli subjektiivista, ja arviointi tapahtui asteikolla yhdestä neljään. Kullekin arvosanalle määritettiin kriteerit, jotka toteutuksen on täytettävä yltääkseen siihen. Taulukko 8 esittää arvioinnissa käytetyn asteikon.

Taulukko 8. Synkronisaation arviointiin käytettävä asteikko.

Arvosana	Vaatimukset arvosanalle
1	Puhe ja leuan liike toimivat epäsynkronisesti. Leuan liike häiritsee puheen ymmärtämistä.
2	Puhe ja leuan liike ovat epäsynkronisoituja suuren osan ajasta. Leuan liike häiritsee puheen ymmärtämistä jossain määrin.
3	Puhe ja leuan liike ovat synkronisoituja suuren osan ajasta. Leuan liike tukee puheen ymmärtämistä jossain määrin.
4	Puhe ja leuan liike ovat täysin synkronisoituja. Leuan liike tukee puheen ymmärtämistä.

Toteutuksen saavuttama arvosana oli kolme. Leuan liikkeen vähäisistä asentovaihtoehdoista huolimatta leuan liike oli suuren osan ajasta synkroninen tuotetun puheen kanssa, ja liike tuki puheen ymmärtämistä. Puheen ja leuan liikkeen hyvällä synkronisaatiolla voidaan hieman kompensoida leuan liikkeen yksinkertaisuutta.

4. JATKOKEHITYS

Toteutuksen aikana heräsi useita ideoita, kuinka työtä voitaisiin parantaa, mutta joiden toteuttamiseen ei ajan tai laitteiden rajallisuuden vuoksi voitu ryhtyä. Parannusten myötä puhesynteesistä, leuan liikkeestä sekä näiden synkronisaatioista olisi mahdollista saada sujuvampaa ja ihmismäisempää.

Servojen ohjauksessa olisi voitu koittaa saada myös lukuoperaatio mukaan ROS-solmuun, jotta tulevaisuudessa leuan servon ääriasentojen määrittäminen onnistuisi helposti. Myös useampia erilaisia suun asentoja olisi voitu implementoida. Puhesynteesin tapauksessa olisi voitu toteuttaa useampia asetuksia, sekä asetuksien vaihtoehtoja kuten tuki uudelle kielelle. Synteesien toteutuksia olisi myös voitu vielä hienosäätää paremmiksi.

4.1. Servon kontrolloiminen

Servon kontrolloimisen osalta suurin ongelma oli se, että lukuoperaatiota varten käytetty kirjasto XL320 ei ollut loppupeleissä yhteensopiva ROS:in omien otsikkotiedostojen kanssa johtuen Arduinon pinnien 1 ja 0 päällekkäisistä määrittelyistä. Se johti osaltaan siihen, että nykyinen servon liikkeen tuottava solmu luottaa suoraan päähän kiinnitetyltä servolta testeissä luettuihin asentojen arvoihin, mikä on sinällään toimiva ratkaisu, sillä leuan servon akseli ei pääse liikkumaan vapaasti päähän kiinnitettynä.

Tulevaisuudessa yhteensopivuusongelma aiheuttaa kuitenkin sen, että jos ja kun leuan liikkeestä vastaava servo vaihdetaan uuteen, joudutaan uuden servon asennot lukemaan, niille on laskettava uudet CRC-tarkistussummat, ja nämä kaikki arvot on vaihdettava servon liikkeen tuottavaan koodiin, mikä tekee koodin uudelleenkäytöstä haastavaa.

Ongelma voitaisiin ratkaista ainakin kolmella seuraavalla tavalla:

- ratkaisemalla Arduinon pinnien 1 ja 0 päällekkäisten määrittelyjen ongelma, jolloin kirjastoa XL320 voitaisiin käyttää ROS-solmussa, mikä helpottaisi servon asentojen lukemista ja kirjoittamista, ja poistaisi tarpeen tarkistussummien laskemiseen itse;
- toteuttamalla perinteinen laitteistotason piiri, ns. ’tri-state buffer’, jolla servon data muunnetaan yksilinjaisesta vuorosuuntaisesta kaksisuuntaiseksi, jolloin kirjastoa XL320 ei tarvita lukuoperaatiota varten, mutta tarkistussummat on laskettava itse; tai
- toteuttamalla asentojen lukemista ja tarkistussummien määrittämistä varten oma erillinen ohjelmansa, joka palauttaa servolta luetut maksimi- ja minimiasennon ja niihin liittyvät tarkistussummat jotka asetetaan ROS-solmun koodiin oikeisiin kohtiin ennen solmun käyttöä.

Nykyisellään ratkaisu, jota servo käyttää on esitellyistä vaihtoehdoista lähimpänä viimeisintä. Servon asennon lukemiselle on toteutettu oma ohjelmansa, joka palauttaa servon silloisen asennon. Kokeilemalla voidaan selvittää leuan ääriasennot ja lukea servon silloiset asennon arvot kokonaislukuina. Kukin luettu asento voidaan sitten antaa parametrina toiselle ohjelmalle, joka muuntaa asennon

kokonaislukumuodosta heksadesimaalimuotoon ja laskee kyseistä asentoa varten tarvittavan käskypaketin tarkistussummat. Ohjelma tulostaa kaikki tarvittavat parametrit Arduinon monitorille, ja kyseiset parametrit voidaan lopuksi vaihtaa servon ROS-solmun koodiin vanhojen arvojen tilalle. Jatkon kannalta järkevin ja toteuttamiskelpoisin vaihtoehto olisi todennäköisesti laitteistotason piirin toteuttaminen piirilevynä, jolloin yhteensopivuusongelman aiheuttavaa kirjastoa ei enää tarvita lukuoperaatiota varten.

4.2. Puhesynteesi

Puhesynteesin asetuksiin voitaisiin implementoida lisää toteutuksia, esimerkiksi kasvattaa kielien lukumäärää. Myös prosodisten ominaisuuksien lisääminen puhesynteesiin toisi lisää ihmismäisyyttä puheeseen, sillä monotoninen puhe ei sovellu kaikkiin mahdollisiin käyttötapauksiin, ja saattaa aiheuttaa ’uncanny valley’ -ilmiötä. Erilaisia puhujia voisi lisätä nauhoituksiin, ja näin voisi lisätä ’sukupuoli’-asetuksiin uusia valintoja. Asetuksen nimen voisi myös muuttaa ’puhujaksi’, jolloin synteesiin voisi lisätä vaihtoehdoiksi eri ihmisten nauhotteita, ja näin ollen äänen kustomisointimahdollisuudet nousisivat entisestään. Uusien asetusten toteuttaminen lisäisi myös puhesynteesin mahdollisia käyttötarkoituksia. Eräs uusi lisättävä asetus voisi olla esimerkiksi kieliopin asetukset.

Vaihtoehtoiseksi puhesynteesiksi voisi lisätä tilastollisen parametrisynteesin. Nykyinen konkatenaatiosynteesin puhe ei ole kaikkein sulavinta, joten synteesien parempi äänen sulavuus voisi nostaa puhutun äänen mielekkyyttä. Lisäksi formanttisynteesissä tuotettu hyvin epäluonnollinen puhe luo robottimaisen äänen. Tilastollisella parametrisynteesillä voitaisiin luoda puhetta, joka on luontevaa ja ihmismäisempää, ja jossa prosodian muuttaminen olisi helpompaa kuin konkatenaatiosynteesissä. Eri toteutuksia voisi näin ollen vaihdella tilanteen mukaan.

Konkatenaatiosynteesin sulavuutta ja ymmärrettävyyttä voisi myös parantaa jatkossa. Nykyinen toteutus aiheuttaa joitain epämiellyttäviä hypähdyksiä äänteiden välissä. Näitä voidaan minimoida käyttämällä tasausta tai luomalla äänitteitä joissa näitä hypähdyksiä ei synny ollenkaan. Äänitteiden parannusta on jo toteutettu puhesynteesiä tehdessä, mutta jotkin äänteet aiheuttavat edelleen pieniä hypähdyksiä, ja näin ollen parantamisen varaa olisi.

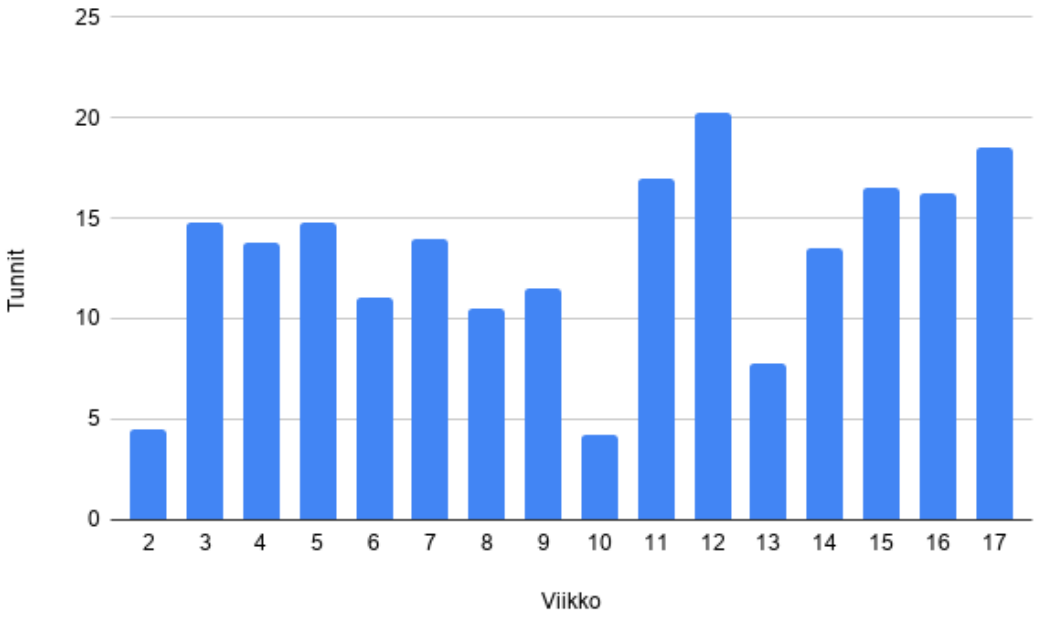
Myös formanttisynteesiä voisi parantaa, sillä nykyinen toteutus on ymmärrettävissä, mutta melko epäselvä. Toteutusta voisi parantaa etsimällä formanteille ja kaistanleveyksille parempia parametrejä. Myös erilaiset aaltomuodot kolmioaallon lisäksi voisivat muuttaa puhetta, mutta on muistettava, että osalla aaltomuodoista voidaan menettää informaatiota.

Käyttökokemuksen parantamiseksi voisi toteuttaa kuuntelijasta riippuvaisen puhesynteesin. Tämä toteutus vaatisi sen, että robotti voisi profiloida henkilön, jonka näkee ja valita asetukset automaattisesti tämän mukaan. Esimerkiksi jos robotti havaitsee edessään naisen, robotti vaihtaisi oman puheensa naisen puheeksi, mutta vain jos se ei ole puhunut jo aikaisemmin näkemälleen ihmiselle, sillä äänen yhtäkkinen muutos tekisi robotille puhumisesta epämukavaa. Lisäksi jos robotista tehdään keskustelijarobotti, voisi robotti valita sanojaan ihmisen sanaston mukaan, jos vain mahdollista.

5. PROJEKTIN KUVAUS

Alustava työnjako oli, että työtehtävät jaettiin kahteen osa-alueeseen, puheeseen ja puhesynteesiin sekä leuan toimintaan ja servon ohjaamiseen, joista molemmat ottivat toisen vastuulleen. Kurssin alussa työskentely oli painottunut kirjoittamiseen ja käytännön työskentely aloitettiin kurssin neljännellä viikolla. Työ oli pääosin omatoimista, mutta tavoitteena oli vähintään kerran viikossa kokoontua työskentelemään yhdessä.

Keskimäärin työhön käytettiin 13.6 tuntia viikossa, ja koko projektin aikana henkilöä kohden kertyi noin 208 tuntia. Työajasta suurin osa kului kurssin alussa lähteiden etsimiseen ja taustakappaleiden kirjoittamiseen, ja kurssin loppupuolella käytännön toteutukseen ja toteutuksesta kirjoittamiseen. Kuva 15 esittää työtuntien keskimääräisen jakautumisen viikoittain.



Kuva 15. Työtuntien keskimääräinen jakautuminen viikoittain henkilöä kohden.

6. YHTEENVETO

Projektin aikana toteutettiin onnistuneesti puhesynteesi ja leuan liike InMoov-robotille. Leuan liikkeen toteutuksessa käytettiin Arduino UNO:a sekä Dynamixelin XL-320-servoa. Puhesynteesinä toteutettiin Python-ohjelmointikielellä konkatenaatiosynteesi, joka toimii sekä mies- että naisäänellä, sekä formanttisynteesi.

Molemmista komponenteista toteutettiin ROS-moduulit, jotka kommunikoivat keskenään ROS:in julkaisija/tilaaja -mallin mukaisesti. Puhesynteesistä vastaava ROS-solmu julkaisee aihekanavalle parametreja, jotka kertovat servon liikkeen toteuttavalle solmulle halutun suun asennon jokaista puhesynteesin syötteeksi annettua kirjainta kohden. Näin saadaan aikaan puheen ja leuan liikkeen synkronisaatio.

Työn tuloksia arvioitiin subjektiivisesti numeerisilla arviointiasteikoilla. Konkatenaatiosynteesi saavutti arvosanan 3/5, mikä tarkoittaa, että puheen välittämä viesti on ymmärrettävissä, formanttisynteesi arvosanan 2/5, mikä tarkoittaa, että puheen välittämä viesti jää epäselväksi, ja synkronisaatio arvosanan 3/4, mikä tarkoittaa, että synkronisaatio on suurimman osan ajasta hyvä. Projektin perusteella voidaan todeta, että erilaisia puhesynteesejä sekä yksinkertainen robotin leuan liikkeen ja puheen synkronisaatio voidaan toteuttaa ROS-ympäristöä hyödyntäen.

Toteutusta voidaan hyödyntää esimerkiksi ihminen-robotti-vuorovaikutuksen tutkimiseen sekä erilaisten havaintoesitysten toteuttamiseen. Jatkokehitystä voisi tehdä puhesynteesin selkeyden ja luonnollisuuden, sekä leuan liikkeen monipuolistamisen osalta.

7. VIITTEET

[1] Cid F., Moreno J., Bustos P. & Núnez P. (2014) Muecas: a multi-sensor robotic head for affective human robot interaction and imitation. Sensors 14, ss. 7711–7737.

[2] Thanh V.N. (2017) A Study of Cerebellum-Like Spiking Neural Networks for the Prosody Generation of Robotic Speech. väitöskirja, School of Engineering, Kagawa University.

[3] Wang Y. & Zhu J. (2016) Artificial muscles for jaw movements. Extreme Mechanics Letters 6, ss. 88 – 95. URL: http://www.sciencedirect.com/science/article/pii/S2352431615300298.

[4] Daumas B., Xu W. & Bronlund J. (2005) Jaw mechanism modeling and simulation. Mechanism and Machine Theory 40, ss. 821 – 833. URL: http://www.sciencedirect.com/science/article/pii/S0094114X05000248.

[5] Hannam A.G. & McMillan A.S. (1994) Internal organization in the human jaw muscles. Critical Reviews in Oral Biology & Medicine 5, ss. 55–89.

[6] Castro-Gonzalez A., Alcocer-Luna J., Malfaz M., Alonso-Martin F. & Salichs M. (2018) Evaluation of artificial mouths in social robots. IEEE Transactions on Human-Machine Systems 48, ss. 369–379.

[7] Lin C.Y., Cheng L.C. & Shen L.C. (2013) Oral mechanism design on face robot for lip-synchronized speech. Teoksessa: 2013 IEEE International Conference on Robotics and Automation, IEEE, ss. 4316–4321.

[8] McGinn C. (2019) Why do robots need a head? the role of social interfaces on service robots. International Journal of Social Robotics URL: https://doi.org/10.1007/s12369-019-00564-5.

[9] Al Moubayed S., Beskow J., Skantze G. & Granström B. (2012) Furhat: A back-projected human-like robot head for multiparty human-machine interaction. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7403 LNCS, ss. 114–130. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84870382387&doi=10.1007%2f978-3-642-34584-5_9&partnerID=40&md5=3917259286b5aac451ca5c3cf1fa4d42, cited By 83.

[10] Kędzierski J., Muszyński R., Zoll C., Oleksy A. & Frontkiewicz M. (2013) Emys—emotive head of a social robot. International Journal of Social Robotics 5, ss. 237–249. URL: https://doi.org/10.1007/s12369-013-0183-1.

[11] Wu T., Butko N.J., Ruvulo P., Bartlett M.S. & Movellan J.R. (2009) Learning to make facial expressions. Teoksessa: 2009 IEEE 8th International Conference on Development and Learning, ss. 1–6.

[12] Zhu X. (2015) Phonetics, articulatory. Teoksessa: J.D. Wright (toim.) International Encyclopedia of the Social & Behavioral Sciences (Second Edition), Elsevier, Oxford, second edition p., ss. 65 – 74.

[13] Raitio T. (2008) Hidden markov model based finnish text-to-speech system utilizing glottal inverse filtering. Master's thesis, Helsinki University of Technology .

[14] Fillebrown T. (1911) Resonance in singing and speaking. Boston: O. Ditson Company; New York: CH Ditson; Chicago: Lyon & Healy.

[15] Ladefoged P. & Johnson K. (2011) A course in phonetics 6th edition. Boston: Thomson Wadsworth .

[16] Suomi K. T.J. & Ylitalo R. (2008) Finnish sound structure : phonetics, phonology, phonotactics and prosody. Oulun yliopiston kirjasto, Oulu. URL: http://urn.fi/urn:isbn:9789514289842.

[17] Vipula & Atula (2018) Human Anatomy and Physiology : For Undergradutae Students of Pharmacy, Nursing, Physiotherapy and Other Paramedical Sciences., nide First edition. Laxmi Publications Pvt Ltd. URL: http://pc124152.oulu.fi:8080/login?url=.

[18] Asheber W.T., Lin C.Y. & Yen S.H. (2016) Humanoid head face mechanism with expandable facial expressions. International Journal of Advanced Robotic Systems 13, s. 29. URL: https://doi.org/10.5772/62181.

[19] URL: http://changingminds.org/explanations/emotions/basic%20emotions.htm.

[20] Story B. (2019) History of speech synthesis. Taylor and Francis, 9-33 s.

[21] Sondhi M. & Schroeter J. (1987) A hybrid time-frequency domain articulatory speech synthesizer. IEEE Transactions on Acoustics, Speech, and Signal Processing 35, ss. 955–967. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0023165217&doi=10.1109%2fTASSP.1987.1165240&partnerID=40&md5=20c762d3bce3d80f93322a83b3147799, cited By 176.

[22] Bawab Z.A., Raj B. & Stern R.M. (2008) Analysis-by-synthesis features for speech recognition. Teoksessa: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, ss. 4185–4188.

[23] Tokuda K., Nankaku Y., Toda T., Zen H., Yamagishi J. & Oura K. (2013) Speech synthesis based on hidden markov models. Proceedings of the IEEE 101, ss. 1234–1252.

[24] Wang Y., Skerry-Ryan R., Stanton D., Wu Y., Weiss R.J., Jaitly N., Yang Z., Xiao Y., Chen Z., Bengio S. et al. (2017) Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 .

[25] Sotelo J., Mehri S., Kumar K., Santos J.F., Kastner K., Courville A. & Bengio Y. (2017) Char2wav: End-to-end speech synthesis. International Conference on Learning Representations 2017 .

[26] Zen H., Tokuda K. & Black A.W. (2009) Statistical parametric speech synthesis. Speech Communication 51, ss. 1039 – 1064. URL: http://www.sciencedirect.com/science/article/pii/S0167639309000648.

[27] Kozima H., Michalowski M. & Nakagawa C. (2009) Keepon: A playful robot for research, therapy, and entertainment. International Journal of Social Robotics 1, ss. 3–18. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84857520097&doi=10.1007%2fs12369-008-0009-8&partnerID=40&md5=6e64d29da70f438d47cf69466d6a8f89, cited By 241.

[28] Pandey A.K. & Gelin R. (2018) A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. IEEE Robotics Automation Magazine 25, ss. 40–48.

[29] Breazeal C.L. (2002) Designing Sociable Robots. Intelligent Robots and Autonomous Agents, A Bradford Book. URL: http://pc124152.oulu.fi:8080/login?url=.

[30] Oh J., Hanson D., Kim W., Han Y., Kim J. & Park I. (2006) Design of android type humanoid robot albert hubo. Teoksessa: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, ss. 1428–1433.

[31] Lee M.K., Forlizzi J., Rybski P.E., Crabbe F., Chung W., Finkle J., Glaser E. & Kiesler S. (2009) The snackbot: Documenting the design of a robot for long-term human-robot interaction. Teoksessa: 2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI), ss. 7–14.

[32] Milliez G. (2018) Buddy: A companion robot for the whole family. Teoksessa: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI '18, Association for Computing Machinery, New York, NY, USA, s. 40. URL: https://doi-org.pc124152.oulu.fi:9443/10.1145/3173386.3177839.

[33] Mäkäräinen M., Kätsyri J. & Takala T. (2014) Exaggerating facial expressions: A way to intensify emotion or a way to the uncanny valley? Cognitive Computation 6, ss. 708–721. URL: https://doi.org/10.1007/s12559-014-9273-0.

[34] Sharkey A.J.C. (2016) Should we welcome robot teachers? Ethics and Information Technology 18, ss. 283–297. Copyright - Ethics and Information Technology is a copyright of Springer, 2016; Document feature - ; Last updated - 2016-12-01.

[35] Edlund J., Gustafson J., Heldner M. & Hjalmarsson A. (2008) Towards human-like spoken dialogue systems. Speech Communication 50, ss. 630 – 645. Evaluating new methods and models for advanced speech-based interactive systems.

[36] Cao H.L., Jensen L., Nghiem X., Vu H., De Beir A., Esteban P., Van de Perre G., Lefeber D. & Vanderborght B. (2019) Dualkeepon: a human–robot interaction testbed to study linguistic features of speech. Intelligent Service Robotics 12, ss. 45–54.

[37] S.m. Z.Q., Wang Z. & Ihsan-ul-haq (2006) Human likeness of humanoid robots exploring the uncanny valley. Teoksessa: 2006 International Conference on Emerging Technologies, ss. 650–656.

[38] Hennig S. & Chellali R. (2012) Expressive synthetic voices: Considerations for human robot interaction. Teoksessa: 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, IEEE, ss. 589–595.

[39] Jonsson I.M. (2009) Social and emotional characteristics of speech-based in-vehicle information systems: impact on attitude and driving behaviour. väitöskirja, Linköping University Electronic Press.

[40] Lee K.M. & Nass C. (2003) Designing social presence of social actors in human computer interaction. Teoksessa: Proceedings of the SIGCHI conference on Human factors in computing systems, ss. 289–296.

[41] Hui C.J., Jain S. & Watson C.I. (2019) Effects of sentence structure and word complexity on intelligibility in machine-to-human communications. Computer Speech & Language 58, ss. 203 – 215.

[42] Wagner P., Beskow J., Betz S., Edlund J., Gustafson J., Eje Henter G., Le Maguer S., Malisz Z., Székely É., Tånnander C. et al. (2019) Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. Teoksessa: Proceedings of the 10th Speech Synthesis Workshop (SSW10).

[43] Sharkey A. & Sharkey N. (2012) Granny and the robots: ethical issues in robot care for the elderly. Ethics and Information Technology 14, ss. 27–40. URL: https://search.proquest.com/docview/927706576?accountid=13031, copyright - Springer Science+Business Media B.V. 2012; Document feature - ; Last updated - 2014-08-30.